Three Essays on Economic Theory

by

Tangren Feng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2020

Doctoral Committee:

Professor Tilman Börgers, Co-Chair
Assistant Professor Shaowei Ke, Co-Chair
Assistant Professor Heng Liu
Professor Scott Page

Tangren Feng

tangren@umich.edu

ORCID iD: 0000-0001-5892-612X

To the three most important females in my life:
Mingjuan Niu, Caicai Chen, and Yichen Feng.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support and inspiration of a number of wonderful individuals. I appreciate all of them for being part of this journey.

First and foremost, I am deeply indebted to Tilman Börgers whos was always extremely welcoming and generous with his time. As an advisor, he encouraged me to challenge existing results and inspired me to study economic theory. As a teacher, his instruction knows no background distinction. When I would have doubt about my work, Tilman gave me the strength and confidence to move on. I would not have come this far without his relentless support and unwavering guidance.

I am exceptionally lucky to have Shaowei Ke as my advisor, collaborator, and friend. I will always value our shared excitement over our first unexpected discovery, the late-night happy hours after a day's hard work, and the joy of eating Gan Bian Fei Chang when the wives were not presented –these are among my fondest memories of the past six years.

I would also like to extend my sincere gratitude to the amazing theory group at the University of Michigan which has been a continual source of encouragement and sound advice. I am especially grateful to Heng Liu, Scott Page, and David Miller. They have been critical to my development as a scholar, and each has greatly helped me to refine my dissertation and presentation skills.

I have been fortunate to be surrounded by a group of close friends: Xinwei Ma, Wenting Song, Ruoyan Sun, Shuqiao Sun, Wenjian Xu, Hang Yu, and Jingyuan Zhai. The wonderful time spent with them have kept Michigan's winter cloud away from my heart. I must also thank Qinggong Wu. The endless discussions with him on (our own and other's) papers and topics of all kinds has been an indispensable part of my academic and personal life.

Finally, my deep gratitude to my family. I am forever indebted to my parents for giving me the opportunities and experiences that have made me who I am. I am deeply grateful to my best friend, great companion, and the love of my life, Caicai Chen, for always being there for me. I thank Caicai and Yichen for making a foreign city home.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

ABSTRACT

This dissertation collects three essays on microeconomic theory.

The first chapter studies a new robustness concept in mechanism design with interdependent values: interim dominant strategy incentive compatibility (IDSIC). It requires truth-telling is an interim dominant strategy for each agent, i.e., conditional on her own private information, the truth-telling maximizes her expected payoff for all possible strategies the other agents could use. In a simple setting with two alternatives and no transfers, we characterize IDSIC together with two other well studied concepts: dominant strategy incentive compatibility (DSIC) and ex post incentive compatibility (EPIC). While both DSIC and EPIC permit only constant mechanisms in sufficiently rich environments, non-constant IDSIC mechanisms exist in any environment. The characterization of IDSIC suggests a simple class of (indirect) binary voting rules: Each agent reports Yes/No. Moreover, if the binary voting rule is also additive, then the indirect mechanism is versatile: It admits an interim dominant strategy equilibrium on all payoff environments and all corresponding type spaces. This chapter is based on the working paper "Robust Binary Voting" (Feng and Wu, 2020).

The most critical issue in evaluating policies and projects that affect generations of individuals is the choice of social discount rate. The second chapter shows that there exist social discount rates such that the planner can simultaneously be (i) an exponential discounting expected utility maximizer; (ii) intergenerationally Pareto—i.e., if all individuals from all generations prefer one policy/project to another, the planner agrees; and (iii) strongly non-dictatorial—i.e., no individual from any generation is ignored. Moreover, to satisfy (i)–(iii), if the time horizon is long enough, it is generically sufficient and necessary for social discounting to be more patient than the most patient individual's long-run discounting, independent of the social risk attitude. This chapter is based on the paper "Social Discounting and Intergenerational Pareto " (Feng and Ke, 2018).

The third chapter studies a decision maker DM who faces a binary choice. DM does not know which alternative is better, but a group of experts do. However, the experts would like DM to make the wrong choice. Given the opposing preferences, is it still possible for

DM to extract useful information from the experts using mechanism design? We answer "Yes": There are mechanisms where truth-telling is a Bayesian or even ex post equilibrium, even though the information leak benefits DM and hurts the expert. On the other hand, if truth-telling is required to be an interim or ex post dominant strategy, then no mechanism extracts information in favor of DM. This chapter is based on the working paper "Getting Information from Your Enemies "(Feng and Wu, 2019).

CHAPTER I

Robust Binary Voting

This chapter studies a new robustness concept in mechanism design with interdependent values: interim dominant strategy incentive compatibility (IDSIC). It requires truth-telling be an interim dominant strategy for each agent, i.e., conditional on her own private information, the truth-telling maximizes her expected payoff for all possible strategies the other agents could use. In a simple setting with two alternatives and no transfers, we characterize IDSIC together with two other well studied concepts: dominant strategy incentive compatibility (DSIC) and ex post incentive compatibility (EPIC). While both DSIC and EPIC permit only constant mechanisms in sufficiently rich environments, non-constant IDSIC mechanisms exist in any environment. The characterization of IDSIC suggests a simple class of (indirect) binary voting rules: Each agent reports Yes/No. Moreover, if the binary voting rule is also additive, then the indirect mechanism is versatile: It admits an interim dominant strategy equilibrium on all payoff environments and all corresponding type spaces.

I.1 Introduction

To vote wisely is not easy. To that end, an agent needs to carefully evaluate the candidates and understand how other people will vote. Such strategic consideration can get complex and exhausting. Therefore, an ideal mechanism would make it easy for agents to determine their unambiguously best vote without having to resort to intricate strategic considerations. In addition to easing agents' cognitive burdens, such a mechanism would also function in a more stable or *robust* fashion, because an agent is likely to adopt that unambiguously best strategy regardless of the many confounding factors she may encounter.

The notion of robustness is captured by the concept of *dominant strategy incentive compatibility (DSIC)* that has been heavily studied in the literature. However, the literature

This chapter is based on the working paper "Robust Binary Voting" (Feng and Wu, 2020).

is based on the keynote that DSIC is too restrictive. Indeed, as the famous Gibbard–Satterthwaite Theorem states, when there are at least three alternatives and preferences are private-value only dictatorships achieve DSIC on unrestricted preference domain (Gibbard (1973) and Satterthwaite (1975)). By this logic, it is then foreseeable that DSIC becomes even more difficult to achieve in the more general interdependent value setting where private information does not pin down one's preference. This we confirm. By studying a model where agents need to collectively choose from *two* alternatives, we observe that, despite there being fewer than enough alternatives to entail the Gibbard–Satterthwaite Theorem, DSIC is nonetheless too restrictive because of preference interdependence. Typically, only constant mechanisms satisfy DSIC, that is, even dictatorships fail the DSIC scrutiny. A non-constant DSIC mechanism exists only if some voters' preferences are *de facto* private-value, and the mechanism is responsive only to these voters.

Many collective choice situations in real life fit the interdependent value setting better, particularly when information about the values of the alternatives is fragmentarily distributed among the population. Classical examples include decision making in committees, legislatures, and juries. It is thus important to understand whether, in the interdependent value setting, there are mechanisms that retain *a reasonable degree of robustness* but are not as austere as ones that satisfy DSIC. This is the main question we address in the paper.

We observe that, in the interdependent value setting, DSIC captures a strong notion of robustness. Given DSIC, an agent can unambiguously determine the best strategy free of any belief about (1) other people's information that directly affects her own preference, and (2) other people's strategies. In other words, DSIC has two orthogonal properties:

1. **Informationally belief-free**: For each agent, there is a strategy that remains a best response given any interim belief about the distribution of the other agents' types, *conditional on equilibrium strategies*.

2. **Strategically belief-free**: For each agent, there is a strategy that remains a best response given any belief about the other agents' strategies, *conditional on her interim belief about the distribution of the other agents' types*.

The informationally belief-free property implies that to come up with a best response, agents do not need to confer with (or even to be conscious of) their own possibly very complicated belief hierarchies. Therefore, indeterminacy in what beliefs the agents would have does not shake the mechanism designer's confidence that they will follow the belief-free optimal strategies, as long as everyone expects everyone else to follow those strategies. In other words, mechanisms that have the informational belief-free property are robust to

misspecification, or the lack of specification, of the agents beliefs about the other agents' types ("informational beliefs").

In contrast, the strategic belief-free property implies that there is an *interim dominant strategy* for every agent that remains a best response regardless of whether she has a correct, or any at all, *strategic* belief about the other agents' strategies. Hence it is easy for an agent to find and take that interim dominant strategy regardless of how she reasons about other people's strategies. Mechanisms that have the strategic belief-free property are robust to the possibility that agents lack adequate strategic sense or understanding ("strategic beliefs").

DSIC mechanisms are exactly the mechanisms that possess both belief-free properties. Having both properties at once certainly makes DSIC mechanisms attractive to a mechanism designer who only has limited or unreliable knowledge about the agents' informational *and* strategic beliefs. The cost is, as we have pointed out, of course the restrictiveness of DSIC. To mitigate this conflict between robustness and leniency, we take the approach of exploring the middle ground by maintaining one belief-free property a time while relaxing the other.

Mechanisms that have the informational belief-free property are, as is well known in the literature (Bergemann and Morris (2005)), ones that satisfy ex post incentive compatibility (EPIC). We develop a characterization for the set of all EPIC (direct revelation) mechanisms in our simple setting with two alternatives and no transfers, and use the characterization to find necessary and sufficient conditions for there to be non-constant EPIC mechanisms. If all agents have the same ex post preference and every agent's preference changes monotonically in her and the other agents' payoff-relevant information in the same fashion, then there exist non-constant EPIC mechanisms.[1] Moreover some of them are ex post Pareto efficient.

Should we be excited about the positive results? To explore how generic environments that admit non-constant EPIC mechanisms might be, we look at continuous type spaces subject to standard regularity conditions. We find that, as long as there is enough preference interdependences and preference heterogeneities so that when indifference curves intersect they do not overlap locally beyond the point of intersection, then every EPIC mechanism must be constant.

Now let us turn to mechanisms that have the strategically belief-free property. We call a mechanism *interim dominant strategy incentive compatible (IDSIC)* if it has the strategic belief-free property, because in such a mechanism every player has an *interim* dominant strategy given her private information. The strategically belief-free property is a popular motivation for DSIC (or strategy-proofness in the literature of voting and market design) mechanism design in complete information or private value settings. However, when it comes to the interdependent value setting, the strategically belief-free property by itself has not

[1]The classical Condorcet Jury model, or common value voting in general, assumes this condition.

received the due attention that we think it deserves.[2] When the strategically belief-free property is discussed in the interdependent values setting, it is discussed as an addendum to DSIC and is hence not analytically separated from the informationally belief-free property. As a result, while DSIC mechanism design prevails, IDSIC mechanism design almost does not exist in the literature. We think ISDIC deserves more attentions than it has. First, as in a complete information or private value setting, a mechanism with the strategically belief-free property allows agents, who might not be strategically sophisticated or "correct" in anticipating an equilibrium, to have an unambiguous optimal strategy and hence behave in a predictable fashion. Second, as we show in the paper, IDSIC is more permissive than DSIC in the interdependent values setting, because non-constant IDSIC mechanisms always exist. This additional permissiveness is desirable for the design of actual mechanisms, especially in cases where there is little ambiguity about the underlying type space and hence the informational belief-free property that comes with DSIC is less important, for instance, when the type space is a common prior type space where the type-generating process is objective and straightforward. Last, as we will also show, even though an agent's interim dominant strategy depends on her type, i.e., her infinite belief hierarchies, in effect only her first-order belief matters. In other words, the agent can determine her interim dominant strategy without being aware of her higher-order beliefs.[3] Therefore, although in an IDSIC mechanism an agent needs to examine her belief to come up with a best response — whereas in a DSIC or EPIC mechanism such examination is not needed at all — this examination is relatively simple.

In the paper, we characterize the set of IDSIC (direct revelation) mechanisms, show that non-constant IDSIC mechanisms always exist, and that they have a very simple structure: For the same agent, all types that have the same strict interim preference regarding the two alternatives are treated equally. The simple structure implies that IDSIC allows a mechanism to dispense with all the bells and whistles that would otherwise be necessary to cater to the rich type space where belief hierarchies could be complicated, because interim preference can be pinned down with just first-order belief. In other words, any IDSIC choice rule can be implemented in a reduced direct mechanism where the agents only report their payoff type and first-order belief. IDSIC mechanisms have a distinct feature that a IDSIC mechanism can only elicits preference rankings (not intensities), intensity still constrains it. Further more, the characterization of IDSIC suggests a class of straightforward indirect mechanisms, which we call *binary additive voting mechanisms*. In a binary additive voting mechanism,

[2]To the extent that there still lack commonly agreed terminologies for the property and the associated incentive compatibility condition. See the Literature subsection for more detailed discussion.

[3]This result applies to more general settings with any finite alternatives.

each agent casts a Yes/No vote. Unlike majority voting mechanisms in which an agent's vote is either pivotal or not, the effect of an agent's vote in binary additive voting mechanisms is independent of the other agents' votes. We show that a binary additive voting mechanism is versatile: it is IDSIC for any type space with respect to any payoff environment.

Literature

There is a massive literature that emphasizes what we mean by robustness, and DSIC is held up as the most thorough. In particular, the literature on robust social choice and voting mostly draws upon the celebrated Gibbard–Satterthwaite impossibility result (Gibbard (1973) and Satterthwaite (1975)) and its extension Hylland (1980), which states that in the private value setting, when there are at least three candidates and any preference profile is possible, then only (random) dictatorship satisfies DSIC.

Follow-up papers look for more positive results in two natural ways. The first one is restricting the preference domain. [4] In contrast, by allowing interdependent preferences, our setting is distinctly different than, and in some cases embeds, the private value setting with unrestricted preference domain that underlies the impossibility results. Moreover, in our setting there are only two candidates, for if there were more then our passage would also be blocked by the impossibility results as long as the private setting with unrestricted domain is embedded in our model, because DSIC, IDSIC and EPIC are equivalent in the private value setting.

The other way of exploring positive results is weakening DSIC in the private value settings. Azevedo and Budish (2019) proposes strategy-proofness in the large (SP-L) which shares the same spirit as IDSIC. SP-L weakens DSIC in two ways. First, it only requires truth-telling be *approximately* optimal *in a large market*. This part is orthogonal to IDSIC. Second, SP-L requires that in the interim stage, truth-telling be best responding to a *subset* (full support, iid distributions) of all possible other agents' reports, rather than best responding to *all* ex post realization of other agents' reports as IDSIC asks. Both SP-L and IDSIC evaluate ICs in the interim stage, but IDSIC is reduced to DSIC with private values and is stronger than SP-L.

There is a growing literature that studies EPIC mechanisms. As far as we know, ours is the first paper that studies EPIC mechanisms for general social choice or voting. Jehiel, Meyer-ter Vehn, Moldovanu and Zame (2006) shows when the payoff state spaces are continuous, agents have interdependent values and multidimensional signals, generically, any deterministic EPIC mechanism (with transfers) is constant. Jehiel et al. is a direct compar-

[4] for example, Moulin (1980), Gershkov, Moldovanu and Shi (2016), and see Barberà (2011) for an excellent survey.

ison with our impossibility result in EPIC. While we do not allow for transfers which makes it harder to find non-constant EPIC mechanisms, we allow for stochastic choice functions and do not require agents' signal be multidimensional. There are several cases that open the door to positive results in Jehiel et al., for example, separable values and one-dimensional signals, neither of which would work in our setting.

The notion of IDSIC is not new (Crémer and McLean (1985)), although it has received far less attention than what we think it deserves. Out of a similar robustness concern, Börgers and Li (2019) proposes a condition, termed as "strategic simplicity", that resembles IDSIC. The main difference is that, under strategic simplicity, an agent has a strategy that is a best response to any scenario that (1) is consistent with her interim belief and (2) where the other agents do not play weakly dominated strategies, whereas, under IDSIC the interim dominant strategy is a best response to any scenario that satisfies (1).[5] Therefore, strategic simplicity is a conceptually weaker condition than IDSIC. Börgers and Li (2019) show that, in private value voting, strategically simple mechanisms are "local dictatorships" in general. There is a large literature on robust mechanism design where robustness is interpreted differently from what we mean in the paper: A mechanism is said to be robust if it is interim incentive compatible[6] with respect to many type spaces à la Harsanyi (1967/1968), or to a very rich type space. There, robustness is interpreted as versatility. In a seminal paper, Bergemann and Morris (2005) show that a mechanism is interim incentive compatible with respect to every type space (or to the universal type space à la Mertens and Zamir (1985)) if and only if it is an EPIC mechanism. Therefore EPIC mechanisms are robust in this sense, too. IDSIC also has a robustness-as-versatility flavor, since an IDSIC mechanism is expected to function well in all situations that share the same type space but differ in the voters' ideas about other people's strategies.

Organization

This chapter is organized as follows. Section I.2 describes the environment and solution concepts. Section I.3 discusses dominant strategy incentive compatibility. Section I.4 presents results on ex post incentive compatibility. Section I.5 considers interim dominant strategy incentive compatibility. Section I.6 concludes. Section I.7 presents minor results on ex post incentive compatibility and Section I.8 collects the proofs.

[5] Although Börgers and Li (2019) focus on the private value setting, the condition can be generalized to the interdependent value setting.

[6] Interim incentive compatibility generalizes Bayesian incentive compatibility to non-common prior type spaces. See Bergemann and Morris (2005).

I.2 Model

Environment

N agents $I = \{1, 2, \ldots, N\}$ need to make a collective choice from two alternatives $a \in A = \{S, R\}$. Every agent receives a payoff of 0 if S — the Status quo or Safe option — is chosen. On the other hand, if R — the Reform or Risky option — is chosen, payoffs to the agents depend on an N-dimensional payoff state of the world $\theta = (\theta_1, ..., \theta_N) \in \Theta_1 \times ... \times \Theta_N = \Theta$. In particular, the payoff to agent i from R being chosen in state θ is expressed as $u_i(\theta)$. In most part of the paper we assume that Θ is finite; however, in one subsection in section I.4 we allow Θ to be infinite. We fix sets I and A through out the paper and call $\langle \Theta, \{u_i\}_{i \in I} \rangle$ the **payoff environment** of the collective choice problem. This payoff environment is common knowledge among the agents.

Information about the true state θ is dispersed among the agents. More specifically, agent i only privately observes θ_i, which we say is agent i's **payoff type**. Agent i also has a (subjective) belief about the other agents' payoff types, and this belief is said to be agent i's first-order belief. Moreover, agent i also has a belief about the other agents' payoff types *and* first-order beliefs, and this belief is said to be agent i's second-order belief. Agent i's higher-order beliefs are defined analogously *ad infinitum*. We call β_i, the agent's (infinite) belief hierarchy, her **belief type**. The agent's payoff type and belief type constitute her **type**.

Types are cumbersome objects to think about and work with because they involve infinite belief hierarchies. Harsanyi (1967/1968) and Mertens and Zamir (1985) show that any type space has a much simpler formulation. Following Bergemann and Morris (2005), the type space is defined as follows. We denote the set of all probability measures on the Borel field of a metric space X by $\Delta(X)$.

Definition I.1. *A type space is a list $\mathcal{T} = \langle T_i, \hat{\theta}_i, \hat{\beta}_i \rangle_{i \in I}$ where for each agent i, T_i is a nonempty finite set of types, and $\hat{\theta}_i, \hat{\beta}_i$ are functions of the form:*

$$\hat{\theta}_i : T_i \to \Theta_i \quad \textit{and} \quad \hat{\beta}_i : T_i \to \Delta(T_{-i})$$

which respectively reflect type t_i's payoff type and belief type.

For each type $t_i \in T_i$ of agent i, $\hat{\theta}_i(t_i)$ is her payoff type and $\hat{\beta}_i(t_i)$ is her belief type. For each payoff type θ_i, there at least exists one type t_i such that $\hat{\theta}_i(t_i) = \theta_i$. We denote $\hat{\beta}_i(t_i)[E]$ the probability that type t_i of agent i assigns to other agents having types t_{-i} in a measurable set $E \subset T_{-i}$.

The infinite belief hierarchy of an agent can be recovered from the simple formulation of types given in Definition I.1. For example, agent i's **first-order belief** function $\hat{b}_i : T_i \to \Delta(\Theta_{-i})$ (which will be of particular importance) can be computed as follows:

$$\hat{b}_i(t_i)[E'] = \sum_{\{t_{-i} | \hat{\theta}_{-i}(t_{-i}) \in E'\}} \hat{\beta}_i(t_i)[t_{-i}],$$

so that $\hat{b}_i(t_i)[E']$ is the probability that type t_i of agent i assigns to the event that the other agents' payoff type profile θ_{-i} is in $E' \subset \Theta_{-i}$.

We fix the type space through out the paper except in subsections I.5.4 and I.5.5.

Mechanisms

We investigate mechanisms by which the agents arrive at a collective decision without the use of side payments. We formulated a mechanism as a messaging game with a choice function as follows.

Definition I.2. *A mechanism is a list $\langle M_1, \cdots, M_N, q \rangle$ such that for each $i \in I$ the set M_i is a nonempty set of messages, and $q : M_1 \times \cdots \times M_N \to [0, 1]$ is a choice function that indicates the probability with which alternative R is chosen.*

Side payments are ruled out because the theoretical objective of this paper is to explore and contribute to the theory of mechanism design with non-transferable utilities. Moreover, not using transfers is a typical constraint to which the design of actual collective choice mechanisms, specifically voting mechanisms, is subject.

On the other hand, we do not require that the mechanism is deterministic. In other words, devices like lotteries can be used in a mechanism such that even if the agents take actions deterministically, the outcome can still be uncertain.

Two classes of mechanisms are of particular interest: The **direct mechanisms** where $M_i = T_i$ for every $i \in I$, and the **fully reduced direct mechanisms** where $M_i = \Theta_i$ for every $i \in I$.[7] Under the direct mechanisms, the agents are asked to report their respective types, whereas under the fully reduced direct mechanism, they are instead asked to report their respective *payoff* types.

Social Choice Rule

A social choice rule $f : T \to [0, 1]$ is a function which maps agents' types to outcomes.

[7] We save the terminology "reduced direct mechanism" for later use.

Solution Concepts

As outlined in the Introduction, we analyze mechanism design subject to three incentive compatibility (IC) conditions:

- Dominant strategy incentive compatibility (DSIC)
- Ex post incentive compatibility (EPIC)
- Interim dominant strategy incentive compatibility (IDSIC)

In the upcoming sections we will formally define these IC conditions. Roughly speaking, DSIC requires that every agent has a (weakly) dominant strategy that maximizes her ex post payoff given any message profile from the other agents in any payoff state θ. [8]

EPIC requires that there is an ex post equilibrium in which every agent's equilibrium strategy maximizes her ex post payoff in any payoff state θ conditional on other agents following their equilibrium strategies. IDSIC requires that every agent has a (weakly) *interim* dominant strategy that maximizes her expected payoff given any message profile from the other agents conditional on her interim belief about the payoff state θ. There is one other IC condition, Bayesian incentive compatibility (BIC), that is popular in the literature. BIC requires that there is a Bayesian Nash equilibrium such that every agent's equilibrium strategy maximizes her expected payoff conditional on her interim belief about the payoff state θ and that other agents follow their equilibrium strategies. We do not discuss BIC because it lacks the robustness properties that we focus on in the paper.

DSIC is the strongest IC condition and BIC is weakest. IDSIC and EPIC are in between, as they arise from respectively relaxing the informational and the strategic belief-free properties from DSIC.

I.3 Dominant Strategy Incentive Compatibility

In this section, we characterizes all DSIC mechanisms and confirm the claim that DSIC is too restrictive in settings with interdependent values.

Dominant strategy incentive compatibility is formally defined as follows:

Definition I.3. *The strategy profile σ^* is a dominant strategy equilibrium of the mechanism $\langle M_1, \cdots, M_N, q\rangle$ if*

$$u_i(\hat{\theta}(t))q(\sigma_i^*(t_i), m_{-i}) \geq u_i(\hat{\theta}(t))q(m_i, m_{-i})$$

[8] We will abuse notations when there are no confusions. In particular, we require the optimal strategies be truth telling when we say a *direct* mechanism is DSIC. The same applies to EPIC and IDSIC.

for all $m \in M$ and all $t \in T$, and all $i \in I$.

That is, for each agent i and type t_i, $\sigma_i^*(t_i)$ maximizes her ex post utility for all possible messages m_{-i} other agents could send and all possible realizations of other agents' types t_{-i}. If a mechanism admits a dominant strategy equilibrium, then it satisfies dominant strategy incentive compatibility.

By the revelation principle, we can focus on truth-telling dominant strategy equilibria of fully reduced direct mechanisms. Hence, without further specifications, all mechanisms in this section refer to fully reduced direct mechanisms.

Definition I.4. *A fully reduced direct mechanism $\langle \Theta_1, \cdots, \Theta_N, q \rangle$ is dominant strategy incentive compatible if*

$$u_i(\theta) q(\theta_i, \theta'_{-i}) \geq u_i(\theta) q(\theta'_i, \theta'_{-i})$$

for all $\theta, \theta' \in \Theta$, and all $i \in I$.

That is, truth telling is a dominant strategy in the ex post stage, i.e., it maximizes each agent's ex post utility for all possible messages θ'_{-i} other agents could send and all possible realizations of other agents' payoff types θ_{-i}.

I.3.1 Characterization

The following definitions will be useful of the characterization.

Definition I.5. *A correspondence $\phi_i : \Theta_i \rightrightarrows \{-1, 0, 1\}$ is an indicator correspondence of agent i if*

$$\phi_i(\theta_i) \ni \begin{cases} 1, & \text{if } u_i(\theta_i, \theta_{-i}) > 0 \text{ for some } \theta_{-i} \in \Theta_{-i} \\ 0, & \text{if } u_i(\theta_i, \theta_{-i}) = 0 \text{ for some } \theta_{-i} \in \Theta_{-i} \\ -1, & \text{if } u_i(\theta_i, \theta_{-i}) < 0 \text{ for some } \theta_{-i} \in \Theta_{-i} \end{cases} . \tag{I.1}$$

$\phi_i(\theta_i)$ contains agent i's possible ex post preferences over $\{S, R\}$ when her private signal is θ_i. For example, $\phi_i(\theta_i) = \{1, -1\}$ means there exists θ_{-i} and θ'_{-i} such that $u_i(\theta_i, \theta_{-i}) > 0$ and $u_i(\theta_i, \theta'_{-i}) < 0$, i.e., given her private signal θ_i, it is possible agent i prefers R over S or prefers R over S in the ex post stage.

The following lemma establishes the link between indicator correspondences and DSIC mechanisms.

Lemma I.1. *q is DSIC if and only if for any $i = 1, ..., N$ and $\theta_i \in \Theta_i$:*

1. *If $1 \in \phi_i(\theta_i)$ then $q(\theta_i, \theta_{-i}) = \max_{\theta'_i \in \Theta_i} q(\theta'_i, \theta_{-i})$ for any $\theta_{-i} \in \Theta_{-i}$.*
2. *If $-1 \in \phi_i(\theta_i)$ then $q(\theta_i, \theta_{-i}) = \min_{\theta'_i \in \Theta_i} q(\theta'_i, \theta_{-i})$ for any $\theta_{-i} \in \Theta_{-i}$.*

An immediate implication of Lemma I.1 is that if there exist one private signal θ_i such that agent i is uncertain about her ex post preferences over $\{S, R\}$ then a DSIC mechanism cannot be responsive to her private signal. Lemma I.2 formalizes the observation.

Definition I.6. *A mechanism $q(\theta_1, \cdots, \theta_N)$ is **responsive** to θ_i if there exist $\theta_i, \theta'_i \in \Theta_i$ and $\theta_{-i} \in \Theta_{-i}$ such that $q(\theta_i, \theta_{-i}) \neq q(\theta'_i, \theta_{-i})$.*

Definition I.7. *Agent i has **quasi-private values** if there does not exist $\theta_i \in \Theta_i$ such that $\{-1, 1\} \subset \phi_i(\theta_i)$.*

An agent who has quasi-private value is certain of whether S is weakly superior to R or R is weakly superior to S based on her private information. In other words, a quasi-private value agent's interim preferences over $\{S, R\}$ are the same as her ex post preferences.

Lemma I.2. *Suppose q is dominant strategy incentive compatible. Then q is responsive to θ_i only if agent i has quasi-private values.*

Lemma I.2 allows us to focus on mechanisms that only heed to agents of essentially private values. Let PI be the set of agents who have quasi-private valuations. We then introduce the binary relation $\Rightarrow_i$ on Θ_i where $i \in PI$: $\theta_i \Rightarrow_i \theta'_i$ if $\phi_i(\theta_i) \ni 1$, $\phi_i(\theta'_i) \ni -1$, or $\theta_i = \theta'_i$. The binary relation captures the dominant strategy incentive compatibilities.

Proposition I.1. *q is DSIC if and only if*

1. *$q(\theta) = q(\theta')$ if $\theta_i = \theta'_i$ for all $i \in PI$;*
2. *$q(\theta) \geq q(\theta')$ if $\theta_i \Rightarrow_i \theta'_i$ for all $i \in PI$.*

The first part of the proposition captures the essence of Lemma I.2: A DSIC mechanism cannot be responsive to any agents other than those have quasi-private values. Also note that if $PI = \emptyset$ then the first part holds for all s and s'. Hence, only constant mechanisms are DSIC. The second part suggests the mechanism needs and only needs to respect quasi-private value agents' ordinal preferences to elicit private information from them. In other words, the mechanism is only informed by ordinal preferences.

According to Proposition I.1, any DSIC choice rule can be indirectly implemented by a mechanism that only collects reports from those who have quasi-private values. Excluding the agents not having quasi-private valuations essentially transforms the environment into a private value setting, and therefore the proposition implies that preference interdependence is in sharp conflict with the existence of non-trivial DSIC mechanisms. In other words, DSIC cannot survive in general interdependent value setting.

Practically, a DSIC mechanism only respects those who have a strong opinion regarding the relative values of R vis-à-vis S, i.e. those who cannot be swayed by additional input from the other agents. In other words, DSIC only allows those who are stubborn to have their voices heard, possibly against objections from all the others agents. Since the purpose of mechanism design is often exactly to rectify the unfortunate situation that would arise in the absence of such a mechanism, DSIC does not achieve much in this respect.

I.3.2 A Partnership Example

Here we use a simple example to illustrate our results. This example will be re-examined when we introduce the other IC conditions.

Two agents need to decide whether they form a partnership or not. Both agents have two possible payoff types, *H(igh)* or *L(ow)*. S represents the joint decision of no partnership, and R represents partnership. The value of partnership depends on the types. Partnership between two high types is very productive, in which case both agents receive a payoff of 4. Partnership between two low types is less productive, in which case both agents receive a payoff of 1. Partnership where types mismatch is counterproductive, in which case both agents receive a payoff of -2.

$u_i(\theta)$	H	L
H	4	-2
L	-2	1

For simplicity, assume that there is a unique belief type associated with each payoff type (and when discussing this example we simply use the payoff type to represent the type of an agent), and the beliefs all come from a common prior such that every state has a probability of 1/4.

Neither agent has quasi-private values, because for either type of the same agent, the sign of her payoff depends on the other agent's type, and hence she is uncertain about her preference ranking regarding S and R despite her private information. It follows from Proposition I.1 that only constant mechanisms are DSIC in this environment. This observation is

striking because there is no conflict of interest between the two agents, yet it is still impossible to robustly (in the DSIC sense) incentivize truthful revelation of private information for non-trivial collective decision making.

I.4 Ex Post Incentive Compatibility

In this section, we first give a characterization of EPIC and two sufficient conditions on the payoff environment for the existence of non-constant EPIC mechanisms. Later, we show that when the payoff type space is continuous and the payoff environment is sufficiently rich, any EPIC mechanism is constant.

Ex post incentive compatibility is formally defined as follows.

Definition I.8. *The strategy profile σ^* is an ex post equilibrium of the mechanism $\langle M_1, \cdots, M_N, q\rangle$ if*

$$u_i(\hat{\theta}(t))q(\sigma^*(t)) \geq u_i(\hat{\theta}(t))q(m_i, \sigma^*_{-i}(t_{-i}))$$

for all $m_i \in M_i$ and all $t \in T$, and all $i \in I$.

That is, for each agent i and type t_i, $\sigma^*_i(t_i)$ maximizes her ex post utility all possible realizations of other agents' types t_{-i} conditional on other agents would play their equilibrium strategies. If a mechanism admits a ex post equilibrium, then it satisfies ex post incentive compatibility.

By the revelation principle, we can again focus on fully reduced direct ex post incentive compatible mechanisms. Again, without further specifications, all mechanisms refer to fully reduced mechanisms in this section.

Definition I.9. *A fully reduced direct mechanism q is ex post incentive compatible if for any $\theta \in \Theta$, $i \in I$ and $\theta'_i \in \Theta_i$,*

$$u_i(\theta)q(\theta_i, \theta_{-i}) \geq u_i(\theta)q(\theta'_i, \theta_{-i}).$$

Therefore, under an EPIC mechanism, truth-telling is a best response for every agent even if the state is common knowledge (given agent i's strategy set is Θ_i not Θ). It equivalently means that truth-telling is a best response regardless of an agent's belief about the distribution of the other agents' signals. In contrast with DSIC, EPIC requires every agent to (correctly) believe that the other agents are truthful.

I.4.1 Characterization

Given a payoff environment $\langle \Theta, \{u_i\}_{i \in I} \rangle$, define a binary relation $\to$ over Θ as follows: $\theta \to \theta'$ if there exists $i \in \{1, ..., N\}$ such that: (1) $\theta_{-i} = \theta'_{-i}$ and (2) $u_i(\theta) > 0$ or $u_i(\theta') < 0$.

If a list of states $(\theta^1, ..., \theta^J)$[9] satisfies $\theta^1 \to \theta^2 ... \to \theta^J$, then this list is called a **path**. Moreover, if $\theta^1 = \theta^J$, then it is a **cycle**. The notation $\theta \rightsquigarrow \theta'$ is used to denote that there is a path from θ to θ', and $\theta \leftrightsquigarrow \theta'$ denotes that there is a cycle from θ to θ'.

It is clear that $\rightsquigarrow$ and $\leftrightsquigarrow$ are both reflexive[10] and transitive, and moreover $\leftrightsquigarrow$ is also symmetric. Since $\leftrightsquigarrow$ is reflexive, symmetric and transitive, it is an equivalence relation. Let C^* denote the equivalence class partition induced by $\leftrightsquigarrow$ on Θ, i.e. θ, θ' are in the same cell in C^* if and only if $\theta \leftrightsquigarrow \theta'$.

It will be useful to abuse notation and extend the $\to$ relation to *sets* of states: For a pair of sets of states c and c', we say $c \to c'$ if there exist $\theta \in c$ and $\theta' \in c'$ such that $\theta \to \theta'$. We extend $\rightsquigarrow$ and $\leftrightsquigarrow$ to sets of states analogously.

A partition C of Θ is said to be **acyclic** if there does not exist distinct $c, c' \in \Theta$ such that $c \leftrightsquigarrow c'$. In one of the proofs we will show that C^* is the finest acyclic partition of Θ

Proposition I.2. *q is EPIC if and only if there are probabilities $\{q_c\}_{c \in C^*}$ such that:*

1. $q(\theta) = q_c$ *if* $\theta \in c$.
2. $q_c \geq q_{c'}$ *if* $c \rightsquigarrow c'$.

Proposition I.2 shows that EPIC mechanisms are mechanisms that respect the $\rightsquigarrow$ relation and the corresponding finest acyclic partition C^* on the state space. If two states θ, θ' are in the same cell of C^*, then EPIC requires the same probability of choosing R in them; if two states are in different cells, then relationship between $q(\theta)$ and $q(\theta')$ needs to agree with the $\rightsquigarrow$ relation. In other words, EPIC mechanisms are only responsive to ordinal information.

The proposition illustrates what EPIC entails in the binary voting environment: Since there are only two alternatives, then in any state an agent either prefers that R be chosen with higher probability or L be chosen with higher probability, modulo indifference. EPIC thus requires that, for any agent, unilaterally deviating to reporting untruthfully weakly reduces the probability of R being chosen in any state where she prefers R, or increases the probability whenever she prefers S. The states are therefore chained by such potential unilateral deviations, and the probabilities of R being chosen must cascade down along the $\rightsquigarrow$ chain to prevent any upstream traffic which represents an untruthful deviation.

[9] The same state is allowed to appear multiple times in the list.

[10] The singleton list $\{\theta\}$ is a (degenerate) path and cycle

I.4.2 Examples

The Partnership Example

Here we come back to the partnership example introduced in previous section. Recall that the payoff matrix is

$u_i(\theta)$	H	L
H	4	-2
L	-2	1

In this example, each state is a cell of the finest acyclic partition C^*, and all ex post incentive compatible mechanisms can be characterized by the following four inequalities implied by the second part of Proposition I.2: the probabilities of choosing R when both agents prefer R, q_1 and q_4, shall be greater than their conterparts, q_2 and q_3, when both agents prefer S.

q	0	1
0	q_1 ≥	q_2
	≥	≤
1	q_3 ≤	q_4

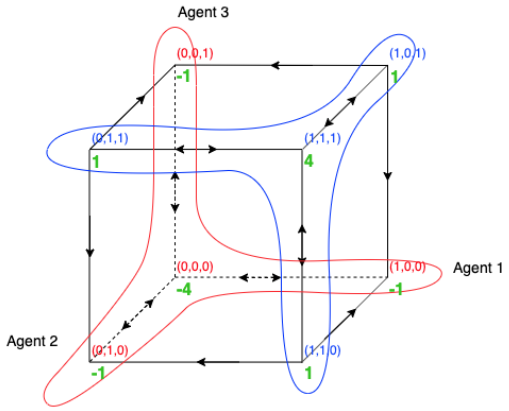
Generalized Condorcet jury

We analyze a classical example of a binary social choice problems—Condorcet jury voting as an example. We consider a generalized version of it.

Each agent each gets a binary signal θ_i taking value of 0 or 1. There exists a payoff function u such that $u_i(\theta) = u(\theta)$ for all i and u is permutation invariant, so the value of R depends on how many 1-signals obtain. Define $\epsilon(\theta) := \sum_{i=1}^{N} \theta_i$. Suppose $u(\theta) \geq u(\theta')$ if and only if $\epsilon(\theta) \geq \epsilon(\theta')$, that is, R is more valuable if there are more 1-signals. This situation generalizes the Condorcet Jury model with the interpretation that the agents are jurors to determine whether to convict or acquit a defendant. A 1-signal is a partial evidence that the defendant is guilty, thus the more 1's the more guilty the defendant could be. S represents the decision to acquit and R to convict. The jurors prefer to acquit if there's not enough guilty evidence or to convict otherwise.

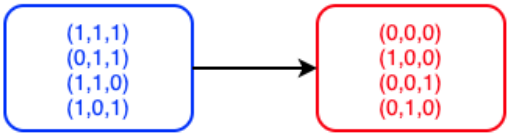
Figure I.1 is an example with $N = 3$. Each vertex of the cube is a state with the corresponding payoff $u(\theta)$ underneath it. For example, the vertex $(0, 0, 0)$ is the state every agent gets signal 0, and the payoff of choosing R in this state is -4 which is below the vertex.

Figure I.1. Condorcet Jury with $N = 3$



The $\rightsquigarrow$ relation over states and the finest acyclic partition C^* and are represented by the arrows and colored sets in the figure. The partition C^* and the $\rightsquigarrow$ relation over sets of states are represented by Figure I.2.

Figure I.2. Finest Acyclic Partition C^*



Hence, a mechanism is EPIC if and only if all the states in the blue cell share the same probability of choosing R, q_1, all the states in the blue cell share the same probability of choosing R, q_2, and $q_1 \geq q_2$.

I.4.3 Existence of Non-constant EPIC Mechanism

We discuss the existence and uniqueness of non-constant EPIC mechanism in this subsection.

Proposition I.3. *The following statements are equivalent:*

1. *There exists a non-constant EPIC mechanism.*

2. *C^* is not a singleton.*

3. Θ can be bipartitioned into c_A and c_B such that $c_B \not\to c_A$.

4. There exist $\theta, \theta' \in \Theta$ such that $\theta \not\to \theta'$.

How useful Proposition I.3 is for a particular environment depends on how easy it is to determine C^*. If the state space is large and has no obvious structure, then determining C^* could be computationally exhausting. Moreover, since the proposition is formulated not directly in terms of the environment parameters, but instead indirectly in terms of the abstract structure underpinning the environment, it is difficult to extract much intuition about when and why a non-constant mechanism exists. To have a better understanding in this respect, we propose two sufficient conditions for the existence of non-constant mechanisms that are formulated directly in terms of the environment parameters.

The first condition is directly inherited from the observation in Proposition I.1 that there are non-constant DSIC mechanisms if and only if there are agents with quasi-private values. Hence, we have the following corollary.

Corollary I.1. *Suppose there exists at least one agent with quasi-private values. Then there exist non-constant EPIC mechanisms.*

Since a DSIC mechanism must also be an EPIC mechanism, the existence of agents with quasi-private values obviously implies that there are non-constant EPIC mechanisms.

The second condition is motivated by the Condorcet jury voting example: It is not unusual that agents might have identical ex post preferences. Formally,

Definition I.10. *Agents have common interests if there exists a payoff function u such that $sgn(u_i(\theta)) = sgn(u(\theta))$ for all $i \in I$ and $\theta \in \Theta$.*

Proposition I.4. *Suppose agents have common interests. Then the ex post Pareto efficient mechanism q^* is EPIC where*

$$q^*(\theta) = \begin{cases} 1, & u(\theta) \geq 0 \\ 0, & u(\theta) < 0 \end{cases}$$

Proposition I.4 indicates common interests is a sufficient condition for the existence of non-constant EPIC mechanisms.

That the Pareto efficient mechanism is EPIC is not surprising given common interests. It is easy to see that the truth-telling strategy profile that simultaneously maximizes every agent's ex post payoff is an ex post equilibrium, because it is impossible for any agent to increase the payoff by unilateral deviation as the upper bound is already reached.

These two sufficient conditions for existence of non-constant EPIC mechanisms — the existence of agents with quasi-private values and common interests — are satisfied in many common environments, yet there are many other environments in which they are violated. Indeed, any general environment in which there is sufficient preference interdependence and preference heterogeneity would violate both conditions. Do non-constant EPIC mechanisms exist in those environments? Yes, there *might* exist non-constant mechanisms. [11] However, as we will show in the extension to continuous payoff state spaces in the next subsection, these two sufficient conditions are "almost necessary" for the existence of non-constant EPIC mechanisms, in the sense that any continuous type space that violates mild generalizations of both conditions only admits constant EPIC mechanisms.

I.4.4 Continuous Type Spaces: A Negative Result

In this subsection, we consider continuous payoff state spaces and show that, apart from mild generalizations of quasi-private values and common interests, all other environments admit only constant EPIC mechanism.

To simplify exposition, we assume $I = \{1, 2\}$, $\Theta = [0, 1]^2$, and u_i is continuously differentiable on Θ.[12]

Let us introduce some useful notation: For agent i, let IC_i denote the set of all payoff states in which i is indifferent between S and R, and let BD_i denote the set of all payoff states θ such that for any $\epsilon > 0$, there is a payoff state where i strictly prefers S to R and there is also a payoff state where i strictly prefers R to S in the $\epsilon-$neighborhood of θ under the Euclidean norm. Clearly $BD_i \subset IC_i$ because u_i is continuously differentiable.

Definition I.11. *Given Θ, the agents' preferences are **generically interdependent** if for any $i \in I$, $\theta \in BD_i$ and $j \neq i$,*

$$\frac{\partial u_i(\theta)}{\partial \theta_j} \neq 0.$$

If preferences are generically interdependent, then when agent i is in a state where she is indifferent between S and R, a slight change in the payoff type of agent j breaks the indifference. In other words, a slight change in j's payoff type matters to i when i is indifferent between S and R. This, of course, cannot hold if i has quasi-private values.

[11] The example illustrated in Figure I.3 in the appendix is a non-constant EPIC mechanism in such an environment.

[12] If non-constant EPIC mechanisms do not exist in a two agent collective choice problem, it does not exist in other collective choice problems.

Definition I.12. *Given Θ, the agents' preferences are* ***generically heterogeneous*** *if for any $\theta \in BD_1 \cap BD_2$ and any $\epsilon > 0$, there exists θ' in the $\epsilon-$neighborhood of θ such that $u_1(\theta')u_2(\theta') < 0$.*

If preferences are generically heterogeneous, then for any payoff state where both agents are indifferent between S and R, there is a close enough payoff state where the agents have opposite preferences. Generic preference heterogeneity says that the two agents cannot agree everywhere: In the least, when both of them agree that S and R are equally good, a slight change in the state can cause them to disagree.

We denote $\Theta \setminus (IC_1 \cup IC_2)$ by $\bar{\Theta}$ which is the set of payoff states in which both agents have strict ex post preferences. Since u is continuously differentiable, the set $IC_1 \cup IC_2$ is of Lebesgue measure zero in $\mathbb{R}^2$.

Proposition I.5. *If the agents' preferences are generically interdependent and generically heterogeneous, then any ex post incentive compatible mechanism is constant over $\bar{\Theta}$.*

Proposition I.5 shows that EPIC can be restrictive at times, and we would like to suggest that this is often the case, as environments where preferences are not generically interdependent and heterogeneous are exceptions rather than the norm. To illustrate the point, suppose in addition that u_1 and u_2 are both strictly increasing in θ. Therefore BD_1 and BD_2 both are curves cutting through the $[0,1]^2$ square. The proposition implies that all EPIC mechanisms are constant as long as there is not a vertical section on the BD_1 curve, there is not a horizontal section on the BD_2 curve, and the two curves overlap only on finitely many points.

Why do non-constant mechanisms fare better on finite payoff sate spaces? One way to understand the reason is thinking about the finite state space as a coarse, low-resolution discretization of the continuous space. For example, instead of being fully conscious of her exact payoff type which can be any real number between 0 and 1, agent $i = 1, 2$ only roughly rounds the her payoff type to the first decimal place, and consequently she in effect has only 10 payoff types: $< 0 \sim 1 >, ..., < 0.9 \sim 1 >$. On the continuous payoff space, such rough rounding implies that the BD_i curve traces along the $t_1 = 0, 0.1, ..., 1$ and $t_2 = 0, 0.1, ..., 1$ grid lines. Obviously this irons the otherwise swerving BD_i curve into horizontal and vertical sections, and moreover the two curves, otherwise not overlapping, are more likely to be squeezed onto the same section of a grid line. Therefore discretization makes the preferences less generically interdependent and less generically heterogeneous, thus admitting non-constant EPIC mechanisms.

I.5 Interim Dominant Strategy Incentive Compatibility

It is useful to first define what interim dominant strategies are.

Definition I.13. *For agent i, strategy σ_i is an interim dominant strategy in mechanism $\langle M_1, \cdots, M_N, q\rangle$ if $U_i(\sigma_i(t_i), \sigma_{-i}|t_i) \geq U_i(m_i, \sigma_{-i}|t_i)$ for any $t_i \in T_i$, $m_i \in M_i$ and $\sigma_{-i} : T_{-i} \to M_{-i}$, where $U_i(\sigma_i, \sigma_{-i}|t_i)$ denotes agent i's interim expected payoff if she follows strategy σ_i, other agents follow strategies given by σ_{-i}, and her type is t_i.*

In plain words, σ_i prescribes for every type t_i of agent i a strategy that maximizes her interim expected payoff regardless of what strategies other agents follow. It is worth noting that agent i's *subjective* belief $\hat{\beta}_i(t_i)$ is used in computing her interim expected payoff.

We can then define interim dominant strategy equilibrium and interim dominant strategy incentive compatibility as follows.

Definition I.14. *Strategy profile σ^* is an interim dominant strategy equilibrium of mechanism $\langle M_1, \cdots, M_N, q\rangle$ if σ_i^* is an interim dominant strategy for every agent $i = 1, ..., N$.*

If a mechanism admits an interim dominant strategy equilibrium, then it satisfies interim dominant strategy incentive compatibility.

As usual, there is a general revelation principle that can simplify analysis by allowing us to focus on direct mechanisms. Later on will discuss a variation of the direct mechanism and respectively develop another revelation principle.

Lemma I.3. *(Revelation Principle 1) Let σ^* be an interim dominant strategy equilibrium of any mechanism $\langle M_1, \cdots, M_N, q\rangle$. Construct a direct mechanism $\langle \overline{M}_1, \overline{M}_2, \ldots, \overline{M}_N, \overline{q}\rangle$ as follows:*

1. *$\overline{M}_i = T_i$ for all $i \in I$.*
2. *$\overline{q}(t) = q(\sigma^*(t))$ for any $t \in T$.*

Then in this direct mechanism, the strategies given by $\overline{\sigma}_i^*(t_i) = t_i$ for all $i \in I$ and $t_i \in T_i$ form an interim dominant strategy equilibrium. Moreover, $\overline{\sigma}^*$ and σ^* are outcome equivalent.

Definition I.15. *A direct mechanism $\langle T_1, ..., T_N, q\rangle$ is interim dominant strategy incentive compatible if*

$$U_i(t_i, \sigma_{-i}|, t_i) \geq U_i(t_i', \sigma_{-i}|, t_i)$$

all $\sigma_{-i} : T_{-i} \to T_{-i}$, for all $t_i, t_i' \in T_i$ and all $i \in I$, where

$$U_i(t_i', \sigma_{-i}|, t_i) = \sum_{t_{-i} \in T_i} \hat{\beta}_i(t_i)[t_{-i}] q(t_i', \sigma_{-i}(t_{-i})) u_i(\hat{\theta}(t_i, t_{-i}))$$

is agent i's expected payoff of reporting t_i' given her type is t_i and other agents' strategy is σ_{-i}.

I.5.1 IDSIC Mechanisms are Higher-Order Belief-Free

An IDSIC mechanism is by definition strategically belief-free, but it is not informationally belief-free, because whether a strategy is interim dominant depends on an agent's belief. In this section, though, we show that an agent's *first-order belief* is sufficient to determine whether a strategy is interim dominant, whereas higher-order beliefs do not matter.

Recall that agent i's first-order belief, when her type is t_i, assigns probability $\hat{b}_i(t_i)[\theta_{-i}]$ to the event that the type profile of the other agents is θ_{-i}.

Definition I.16. *An strategy profile σ is **higher-order belief-independent** if $\sigma_i(t_i) = \sigma_i(t_i')$ for any t_i, t_i' such that $\hat{\theta}_i(t_i) = \hat{\theta}_i(t_i')$ and $\hat{b}_i(t_i) = \hat{b}_i(t_i')$.*

In words, a higher-order belief-independent strategy profile prescribes the same strategy for all types of agent i that have the same payoff type and first-order belief.

We are ready to start showing the connection between IDSIC and the higher-order belief-free property, To set the stage, let us introduce one more notation: Let $IDM_i(t_i)$ [13] denote the set of all messages m_i such that there is some interim dominant strategy σ_i where $\sigma_i(t_i) = m_i$. It is obvious that σ_i is an interim dominant strategy if and only if $\sigma_i(t_i) \in IDM_i(t_i)$ for every $t_i \in T_i$.

Lemma I.4. *$IDM_i(t_i) = IDM_i(t_i')$ for any $i \in I$ and any $t_i, t_i' \in T_i$ such that $\hat{\theta}_i(t_i) = \hat{\theta}_i(t_i')$ and $\hat{b}_i(t_i) = \hat{b}_i(t_i')$.*

The Lemma essentially shows that if two types of agent i has the same payoff type and first-order beliefs, then they are strategically "identical" if we focus on interim dominant strategy equilibria, because the two types are presented with the same set of messages that

[13] *IDM* represents "interim dominant messages"

dominate other messages at the respective interim stages. This observation leads to the following Proposition.

Proposition I.6. *Mechanism $\langle M_1, ..., M_N, q\rangle$ has an interim dominant strategy equilibrium if and only if it has an interim dominant strategy equilibrium that is higher-order belief-independent.*

The Proposition shows that, in addition to the built-in strategically belief-free property, any IDSIC mechanism also has what we may call the "informationally higher-order belief-free" property, as there is always an interim dominant strategy equilibrium that does not depend on higher-order beliefs. In comparison, the informationally belief-free property introduced in the Introduction is, in this sense, informationally higher-order *and* first-order belief-free.

Proposition I.6 and Lemma I.5 suggest that, to find an interim dominant strategy equilibrium, or to find an IDSIC mechanism, it is without loss of generality to focus only on higher-order belief-independent strategy profiles where only payoff types and first-order beliefs matter. Since the payoff types and first-order beliefs type are of a particular importance with respect to IDSIC, it is useful to refer to them with special terminology:

Definition I.17. *Agent i's **reduced type** $h_i := (\theta_i, b_i)$ consist of her payoff type and her first-order belief. Agent i's reduced type space is $H_i := \Theta_i \times B_i$.*

Naturally, there is a Revelation Principle with respect to reduced types.

Lemma I.5. *(Revelation Principle 2) Let σ^* be a higher-order belief-independent interim dominant strategy equilibrium of a mechanism $\langle M_1, M_2, \ldots, M_N, q\rangle$. Construct a reduced direct mechanism $\langle \overline{M}_1, \overline{M}_2, \ldots, \overline{M}_N, \overline{q}\rangle$ as follows:*

1. *$\overline{M}_i = H_i$ for all $i \in I$.*
2. *$\overline{q}(h) = q\big(\sigma^*(h)\big)$ for all $h \in H = H_1 \times ... \times H_N$.*

Then in this reduced direct mechanism the strategies given by $\overline{\sigma}_i^*(t_i) = (\hat{\theta}_i(t_i), \hat{b}_i(t_i))$ for all $i \in I$ and $t_i \in T_i$ form an interim dominant strategy equilibrium that is higher-order belief-independent. Moreover, $\overline{\sigma}^*$ and σ^* are outcome equivalent.

In the reduced direct mechanism as constructed in the Lemma, an agent reports her reduced type, instead of her original type. It may seem that reporting the reduced type $h_i = (\theta_i, b_i)$ seems to be a more tedious job than reporting the original type t_i, as agent i might need

to derive her payoff type and first-order belief from her type t_i, which introduces additional labor. This appearance, though, is merely an artifact of the Harsanyian formulation of the type space. Indeed, the type t_i is an abstraction of an agent's infinite belief hierarchies, which the agent is likely to be conscious of, whereas her payoff type and first-order belief are more concrete and salient objects and hence are more likely to be in her awareness and more easily to be explicitly reported.

Remark I.1. *All results in this subsection are not restricted to the special two-alternative setting of the paper, as they also apply to settings with any finite number of alternatives.*

I.5.2 Characterization

In this subsection we characterize the set of all IDSIC reduced direct mechanisms.

For any agent i and $h_i = (\theta_i, b_i) \in H_i$ define

$$\overline{\alpha}_i(\theta_i, b_i) := \sum_{\{\theta_{-i} | u_i(\theta_i, \theta_{-i}) > 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}),$$

$$\underline{\alpha}_i(\theta_i, b_i) := \sum_{\{\theta_{-i} | u_i(\theta_i, \theta_{-i}) < 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}),$$

and

$$\alpha_i(\theta_i, b_i) := \sum_{\theta_{-i} \in \Theta_{-i}} b_i(\theta_i) u_i(\theta_i, \theta_{-i}).$$

Clearly we have $\underline{\alpha}_i(h_i) \leq 0 \leq \overline{\alpha}_i(h_i)$ and $\underline{\alpha}_i(h_i) + \overline{\alpha}_i(h_i) = \alpha_i(h_i)$. The sign of $\alpha_i(h_i)$ is positive/negative when agent i at the interim stage prefers R/S. $\overline{\alpha}_i(h_i)$ is the expected payoff to agent i of reduced type h_i from the choice rule that chooses R whenever R is ex post preferred to S. Similarly $\underline{\alpha}_i(h_i)$ is the expected payoff from the choice rule that chooses R whenever S is ex post preferred to R.

Define $H_i^+ := \{h_i \in H_i | \alpha_i(h_i) > 0\}$, $H_i^- := \{h_i \in H_i | \alpha_i(h_i) < 0\}$, and $H_i^0 := \{h_i \in H_i | \alpha_i(h_i) = 0\}$ for each $i \in I$. H_i^+ is the set of reduced types in which agent i strictly prefers R over S in the interim stage, H_i^- is the set of reduced types in which agent i strictly prefers S over R in the interim stage, and H_i^0 is the set of reduced types in which agent i is indifferent.

The following lemma will be useful for us to develop a characterization of IDSIC reduced revelation mechanisms.

Lemma I.6. *q is IDSIC if and only if for any $i = 1, ..., N$, $h_i, h_i' \in H_i$, and $h_{-i}, h_{-i}' \in H_{-i}$:*

$$\underline{\alpha}_i(h_i)\big(q(h_i, h_{-i}) - q(h_i', h_{-i})\big) + \overline{\alpha}_i(h_i)\big(q(h_i, h_{-i}') - q(h_i', h_{-i}')\big) \geq 0. \tag{I.2}$$

The Lemma is based on the observation that, among many possible incentive constraints, the only one that is binding for agent i of reduced type h_i corresponds to the case that the other agents coordinate on on the same message profile (h_{-i}) whenever agent i of type h_i ex post prefers S, or they coordinate on another message profile (h_{-i}') whenever agent i of type h_i ex post prefers R.

The characterization of IDSIC reduced direct mechanisms will depend on crucial parameters defined as follows: For agent i and reduced type h_i,

$$\rho_i(h_i) := \begin{cases} \dfrac{\overline{\alpha}_i(h_i)}{-\underline{\alpha}_i(h_i)} & \text{if } h_i \in H_i^+ \\ 1 & \text{if } h_i \in H_i^0 \\ \dfrac{-\underline{\alpha}_i(h_i)}{\overline{\alpha}_i(h_i)} & \text{if } h_i \in H_i^- \end{cases}$$

And for agent i,

$$\rho_i := \min_{h_i \in H_i} \rho_i(h_i).$$

It is easy to verify that $\rho_i \geq 1$.

We first characterize IDSIC reduced direct mechanisms where there are no types indifferent between S and R at the interim stage. The general characterization is given later on at Proposition I.8.

Proposition I.7. *Suppose $H_i^0 = \emptyset$ for all i. A reduced direct mechanism q is IDSIC if and only if for any agent i and $h_i \in H_{-i}$, there are two numbers $\overline{q}_i(h_{-i})$ and $\underline{q}_i(h_{-i})$, where $\overline{q}_i(h_{-i}) \geq \underline{q}_i(h_{-i})$, such that:*

1.

$$q(h_i, h_{-i}) = \begin{cases} \overline{q}_i(h_{-i}) & \text{if } h_i \in H_i^+ \\ \underline{q}_i(h_{-i}) & \text{if } h_i \in H_i^- \end{cases}$$

2.

$$\max_{h_{-i} \in H_{-i}} \Big(\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i})\Big) \leq \rho_i \min_{h_{-i} \in H_{-i}} \Big(\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i})\Big).$$

Condition 1 is noteworthy in that all types with the same interim ordinal preference are treated equally by the mechanism. In other words, an IDSIC reduced mechanism elicits preference rankings but has to ignore preference intensities. When H_i^0 is empty, the interim

preference ranking that agent i can have is at most two, and hence conditional on the other agents' reports h_{-i}, agent i's marginal influence on the choice probability is binary: high (leading to $\overline{q}_i(h_{-i})$) or low (leading to $\underline{q}_i(h_{-i})$). It is then easy to see that the mechanism can be simulated by a much simpler mechanism, which we call a **binary voting mechanism**. In a binary voting mechanism, every agent is given two messages: R and S. The R message is interpreted as a vote supporting the alternative R, whereas the S message is a vote supporting the alternative S. A unilateral switch from sending message S to sending R raises the chance that alternative R is chosen. Binary voting is clearly the most common and important voting format used for bi-candidate public choice.

Condition 2 shows, despite that an IDSIC reduced direct mechanism does not elicit an agent's interim preference intensity, that intensity still constrains the mechanism. To see this, observe that when the other agents report h_{-i}, agent i's marginal voting power can be represented by $\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i})$, which is the marginal increase in the probability of R being chosen if i unilaterally switches from reporting to be a type in H_i^- to reporting to be a type in H_i^+. Condition 2 says that this marginal voting power cannot fluctuate too widely with h_{-i}, i.e, the ratio of agent i's maximal marginal voting power $\max_{h_{-i}}(\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i}))$ to her minimal marginal voting power $\min_{h_{-i}} \overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i})$ cannot be higher than the parameter ρ_i which depends on the agent's interim preference intensities.

There is a special class of mechanisms such that every agent's marginal voting power is constant, and hence Condition 2 is immediately satisfied. This class of mechanism, which we will formally define as "additive mechanisms" in Definition I.18 and discuss with more details, will turn out to be IDSIC with respect to any type space based on any payoff state space. In other words, additive mechanisms are not only robust, but also versatile.

On the other hand, majority voting mechanisms, where R is chosen if and only if there are more than k votes for R, are not IDSIC exactly because an agent's marginal voting power fluctuates too much. Indeed, when the agent is pivotal, her marginal voting power is the maximal 1, whereas when he is not pivotal, her marginal power is the minimal 0.

Now we drop the assumption that H_i^0 is empty for every i and present the characterization for IDSIC reduced direct mechanisms in this general case.

Proposition I.8. *A reduced direct mechanism q is IDSIC if and only if for any agent i and $h_{-i} \in H_{-i}$ there are two numbers, $\overline{q}_i(h_{-i})$ and $\underline{q}_i(h_{-i})$ where $\overline{q}_i(h_{-i}) \geq \underline{q}_i(h_{-i})$, such that:*

1.

$$q(h_i, h_{-i}) \begin{cases} = \overline{q}_i(h_{-i}) & \text{if } h_i \in H_i^+ \\ \in [\underline{q}_i(h_{-i}), \overline{q}_i(h_{-i})] & \text{if } h_i \in H_i^0 \\ = \underline{q}_i(h_{-i}) & \text{if } h_i \in H_i^- \end{cases}$$

2. If $H_i^0 = \emptyset$, then

$$\max_{h_{-i} \in H_{-i}} \Big(\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i}) \Big) \leq \rho_i \min_{h_{-i} \in H_{-i}} \Big(\overline{q}_i(h_{-i}) - \underline{q}_i(h_{-i}) \Big).$$

If $H_i^0 \neq \emptyset$, then $q(h_i, h_{-i}) - q(h_i', h_{-i})$ is independent of h_{-i}.

The biggest difference between Proposition I.7 and Proposition I.8 is that the existence of an reduced type that is interim indifferent between S and R immediately pushes the parameter ρ_i to the lower bound 1, which implies that agent i's marginal voting power has to be constant.

I.5.3 Example

Here we come back to the partnership example again. Recall that we assume there is a unique belief type associated with each payoff type and the beliefs all come from a common prior such that every state has a probability of 1/4. The payoff matrix is

$u_i(\theta)$	H	L
H	4	-2
L	-2	1

In order to find all IDSIC mechanisms in this example, we first calculate agents' interim preferences. For example, $\alpha_1(H) = \frac{1}{4}4 + \frac{1}{4}(-2) = 1/2$. Then we have $\alpha_2(H) = 1/2 > 0$ and $\alpha_1(L) = \alpha_2(L) = -1/4 < 0$. Monotonicity conditions require $q(H,H) \geq q(L,H), q(H,L) \geq q(L,L), q(H,H) \geq q(H,L), q(L,H) \geq q(L,L)$.

Second, we compute each agent's intensity ratio. When agent 1 is the H type, $\overline{\alpha}_1(H) = \frac{1}{4}4 = 1$ and $\underline{\alpha}_1(H) = \frac{1}{4}(-2) = -1/2$. Then $\rho_1(H) = \overline{\alpha}_1(H)/|\underline{\alpha}_1(H)| = 2$. Similarly, we have $\rho_1(H) = \rho_2(H) = \rho_2(L) = 2$ which leads $\rho_1 = \rho_2 = 2$. Smoothness conditions are the following, $\dfrac{1}{2} \leq \dfrac{q(H,H) - q(L,H)}{q(H,L) - q(L,L)} \leq 2$, $\dfrac{1}{2} \leq \dfrac{q(H,H) - q(H,L)}{q(L,L) - q(L,H)} \leq 2$.

Now suppose $q(H,H) = 1$, $q(H,L) = 1/2$, and $q(L,L) = 0$. Then q is IDSIC if and only if $1/4 \leq q(L,H) \leq 3/4$.

I.5.4 Universal Existence of Non-Constant IDSIC Mechanisms

In this section, we show that non-constant IDSIC mechanisms universally exist over all type spaces. In fact, they are closely related to additive mechanisms that we already informally mentioned in Section I.5.2. Here we formally define them:

Definition I.18. *An indirect mechanism $\langle M_1, ..., M_N, q\rangle$ where $|M_i| \geq 2$ for all i is an **additive mechanism** if there exist functions $\pi_i^q : M_i \to [0,1]$ such that $q(m_1, ..., m_N) = \sum_{i\in I} \pi_i^q(m_i)$.*

In an additive mechanism, every message m_i has a "score" $\pi_i^q(m_i)$ attached to it, and the eventual probability that R is chosen is the sum of the scores of the chosen messages.

Proposition I.9. *Fix a payoff environment. A mechanism $(M_1, ..., M_N, q)$ is IDSIC in all type spaces if and only if it is additive.*[14]

The Proposition establishes the universal existence of non-constant IDSIC mechanisms, because additive mechanisms are always IDSIC, and because most additive mechanisms are non-constant. Moreover, it also shows that additive mechanisms are versatile, in the sense that they are IDSIC with respect too many (in this case, all) type spaces, and hence are expected to function well even if the underlying type space is uncertain. We will discuss this versatility aspect in the next subsection.

Additive mechanisms also have a close relation with a class of well-known social choice rules: random dictatorships.

Definition I.19. *A social choice rule $q : T \to [0,1]$ is a **random dictatorship** if there exist numbers $(\lambda_i^q)_{i\in I}$, where $\lambda_i^q \in [0,1]$ and $\sum_{i\in I} \lambda_i \leq 1$, and functions $\mu_i^q : T_i \to [0,1]$ where $\mu_i^q(t_i) = 1$ if $\alpha_i(t_i) > 0$ and $\mu_i^q(t_i) = 0$ if $\alpha_i(t_i) < 0$, and a constant $\tilde{\lambda}^q \in [0, 1 - \sum_{i\in I} \lambda_i]$, such that*

$$q(t) = \sum_{i\in I} \lambda_i^q \mu_i^q(t_i) + \tilde{\lambda}^q.$$

Under a random dictatorship rule, agent i has chance (with a probability of λ_i^q) to be the "dictator" in the event of which her interim preferred alternative is chosen by the rule. There is also a chance (with a probability of $\tilde{\lambda}^1$) that no agent is chosen to be the random dictator and R is in this case chosen with certainty.

[14] We rule out mechanisms with trivial message sets $|M_i| = 1$ for some i in the proposition.

The following two result establish the close connection between additive mechanisms and random dictatorship rules.

Proposition I.10. *Suppose mechanism $(M_1, ..., M_N, q)$ is an additive mechanism, and σ^* is an interim dominant strategy equilibrium of it. Then the social choice function $q \circ \sigma^* : T \to [0, 1]$ is a random dictatorship.*

Proposition I.11. *If q is a random dictatorship, then the direct mechanism (T, q) is an additive voting mechanisms.*

How should we interpret the results? Do they mean that IDSIC mechanisms in general are not ideal, because they are "dictatorial" by nature? We recommend that "dictatorship" not be taken at the face value, because random dictatorships are social choice rules, not mechanisms. In other words, the actual mechanism that induces a random dictatorship need not entail the seemingly undemocratic element of someone dictating a decision, possibly against the prevailing public opinion. Proposition I.10 shows that additive mechanisms, which can be perfectly democratic and fair *formally and actually* if every agent has the same marginal voting power π_i^q, induces a random dictatorship rule nonetheless. We observe that the random "dictator" under a random dictatorship is as dictatorial as the median voter in the Median Voter Theorem who *de facto* decides the voting outcome, which is to say not very much dictatorial at all.

I.5.5 Versatile IDSIC Mechanisms

In this sub-section we show that there are IDSIC mechanisms that have a nice property: versatility. A mechanism is versatile if it is able to handle a wide variety of situations and is and less reliant on the configuration of the environment. "Aye-Nay" voting used in many legislative procedures, for instance, is versatile, as it handles a budgeting bill as well as an impeachment motion, despite the great difference between these two issues. It is no wonder why such a mechanism has been consistently used over history and across the globe.

Now we set the stage to formally discuss versatility. We identify three building blocks of any collective choice environment: The underlying payoff state space Θ, the payoff functions $\{u_i\}_{i \in I}$ defined with respect to Θ, and the type space T. We jointly call these three elements the **environment** concerning the collective choice problem, and denote it as E. Let $\mathscr{E}$ denote the set of all environments, where, specifically, the underlying payoff state space may vary.

Definition I.20. *A mechanism is **versatile** with respect to a given incentive compatibility if the mechanism satisfies that incentive compatibility given any $E \in \mathscr{E}$.*

Our discussion on DSIC and EPIC shows that no non-constant mechanisms are versatile with respect to those two incentive compatibility conditions. However, non-constant mechanisms that are versatile with respect to IDSIC do exist. In fact, as the following result shows, they have a very simple characterization.

Proposition I.12. *Any binary additive voting rule is versatile with respect to IDSIC.*

Proposition I.12 shows that even if the designer is agnostic not only to the belief environment (type spaces) but also to the payoff environment, she is able to find a mechanism that is IDSIC.

Moreover, when we know more about the environment, we can find more mechanisms other than binary additive voting mechanisms that are IDSIC.

Fix any type space, and an agent i and her type t_i. Let $v_i(t_i) := \dfrac{\overline{\alpha}_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i))}{|\underline{\alpha}_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i))|}$ be called this type's **virtual type**. For any environment E let $V := V_1 \times ...V_N$ where $V_i = \{v_i(t_i) : t_i \in T_i\}$ denotes the **virtual type space** induced by E. Note that the virtual type space of any original type space is a subspace of $(\mathbb{R} \cup \{-\infty, \infty\})^N$. Given any environment, an agent can compute her virtual value (which is a real number) by considering her payoff type and first-order belief.

Proposition I.13. *Given a binary voting mechanism q. If an environment E induces a virtual type space V such that $v_i \notin (1/\eta_i, \eta_i)$ for all $v_i \in V_i$ and $i \in I$, then q is IDSIC on E, where*

$$\eta_i = \frac{\max_{m_{-i}} q(1, m_{-i}) - q(0, m_{-i})}{\min_{m_{-i}} q(1, m_{-i}) - q(0, m_{-i})}$$

for all $i \in I$.

If the designer knows more about the environment, the mechanism may deviate more from binary additive voting without violating IDSIC.

I.5.6 *DSIC = EPIC ∩ IDSIC*

In this subsection, we show that the joint of the two "qualified" robustness conditions — IDSIC and EPIC — is exactly DSIC.

Proposition I.14. *A mechanism is DSIC if and only if it is EPIC and IDSIC.*

The "only if" direction is immediate. For better exposition, we assume the type space is the payoff type space. In order to understand the intuition behind the "if" direction, we shall first recall the defining property of DSIC is that it can not be responsive to any not privately informed agents. Consider a not privately informed agent i. Suppose agent i's interim preferences is R, $\alpha_i(\theta_i) > 0$, upon observing her own signal θ_i, IDSIC requires $\theta_i \in \text{argmax}_{\theta_i'} q(\theta_i', \theta_{-i})$ for all θ_{-i}. Since agent i doesn't have quasi private value, there exists θ_{-i} such that $u_i(\theta_i, \theta_{-i}) < 0$ in which her ex post preferences is L. EPIC demands $\theta_i \in \text{argmin}_{\theta_i'} q(\theta_i', \theta_{-i})$ for that θ_{-i}. As a result, $q(\theta_i, \theta_{-i})$ is a constant function over θ_i.

In other words, EPIC respects agents' ex post preferences while IDSIC follows agents' interim preferences. Disagreements between these two arise whenever the agent does not have quasi private values. The tension is so severe that complying with one leads to a violation of the other. Consequently, any EPIC and IDSIC mechanism cannot be responsive to any agent who does not have quasi private value.

I.6 Conclusion

We study robust mechanisms without transfers in a setting where there are two alternatives, the agents' preferences are interdependent, and the underlying type space is rich. Three notions of robustness (incentive compatibilities) are examined. The first two are widely used: dominant strategy incentive compatibility and ex post incentive compatibility. The former permits nothing but constant mechanisms and the latter is more lenient, but only so in finite type spaces — when the type space is continuous, non-constant mechanisms are again ruled out under weak conditions.

The interdependent values setting enables us to study another natural notion of robustness— *interim dominant strategy incentive compatibility*. It requires that each agent has a weakly interim dominant strategy— that is, conditional on each agent's own private information, her strategy must maximize her expected payoff for all possible strategies the other agents could use. We establish a revelation principle that allows the mechanisms to simply ask the agents to reveal their payoff type and first-order beliefs. The characterization suggests a simple binary voting rule: Each agent reports Yes/No to the mechanism. Moreover, if the binary voting rule is also additive (each agent's influence is independent with other agents reports), then the indirect mechanism is versatile: It admits interim dominant strategy equilibrium on all payoff environments and all corresponding type spaces.

I.7 Additional Results on EPIC Mechanisms

I.7.1 Monotonicity and Efficiency

It turns out Monotonicity plays a major role on the efficiency property of EPIC mechanisms. We define $\Theta_+ := \{\theta \in \Theta | u_i(\theta) > 0 \text{ for all } i \in I\}$ and $\Theta_- := \{\theta \in \Theta | u_i(\theta) < 0 \text{ for all } i \in I\}$. Θ_+ is the set of states in which all agents strictly prefer R over S, and Θ_- is the set of states in which all agents strictly prefer S over R.

Definition I.21. *A mechanism q is ex post Pareto efficient if $q(\theta) = 1$ for all $\theta \in \Theta_+$ and $q(\theta) = 0$ for all $\theta \in \Theta_-$.*

A mechanism q is ex post efficient if it respects unanimity of strict preference on the part of the agents.

Definition I.22. *A mechanism q is ex post efficient in the range if* $\min_{\theta \in \Theta_+} q(\theta) \geq \max_{\theta \in \Theta_-} q(\theta)$.

A mechanism q is ex post efficient in the range if the smallest probability of choosing R when all agents prefer R is higher than the greatest probability of choosing R when when all agents prefer S.

Definition I.23. *A mechanism q is grossly Pareto inefficient if* $\max_{\theta \in \Theta_+} q(\theta) < \min_{\theta \in \Theta_-} q(\theta)$.

A mechanism q is grossly efficient in the range if the largest probability of choosing R when all agents prefer R is lower than the smallest probability of choosing R when when all agents prefer S.

In any state θ, agent i may be of the following three types in terms of her preference over R vs. S: she may prefer R to S ($\text{sgn}(u_i(\theta)) > 0$), she may be indifferent ($\text{sgn}(u_i(\theta)) = 0$), or she may prefer S to R ($\text{sgn}(u_i(\theta)) < 0$).

Definition I.24. *$u_i(\theta)$ is monotone with respect to θ_j if there is a linear order $\succ_j^{(i)}$ on Θ_j such that $\theta_j \succ_j^{(i)} \theta'_j$ implies*

$$sgn(u_i((\theta_j, \theta_{-j})) \geq sgn(u_i((\theta'_j, \theta_{-j}))$$

for all θ_{-j}.

Monotonicity means that agent i has a clear interpretation of agent j's private signal: higher ranked θ_j implies higher payoff of choosing R, regardless of θ_{-j}. In the case that u_i

is not monotone w.r.t. θ_j, agent i's interpretation of θ_j depends on the realization of other private signals.

Also observe that, according to our definition, every private value utility function is monotone.

We say that the environment or $\{u_i\}_{i\in I}$ is monotone if every agent has monotone preference. Formally,

Definition I.25. *$\{u_i\}_{i\in I}$ is monotone if for each agent i there is a collection of linear orders $\{\succ_j^{(i)}\}_{j\in I}$ on $\{\Theta_j\}_{j\in I}$ such that if $\theta_j \succ_j^{(i)} \theta'_j$, then*

$$sgn(u_i((\theta_j, \theta_{-j})) \geq sgn(u_i((\theta'_j, \theta_{-j}))$$

for all θ_{-j} and all i.

Though monotonicity seems to be strong, it is not strong enough to insure every EPIC mechanism is Pareto efficient. Figure I.3 shows an extreme counterexample. Similar to Figure I.1, each agent has two possible signals $\{0, 1\}$, each vertex of the cube represents a state, and the green vector under each state is the corresponding payoff vector. For example, vertex $(0, 1, 1)$ represents the state agent 1 gets signal 0 and both agent 2 and 3 get signal 1. The payoff vector $(-2, 2, 2)$ means the ex post payoff of choosing R in state $(0, 1, 1)$ is -2 for agent 1, and 2 for agent 2 and agent 3. It can be verified that $\{u_i\}_{i\in I}$ is monotone, but no EPIC mechanism is Pareto efficient. Furthermore, since $\Theta_- = \{(1, 1, 1)\}$, $\Theta_+ = \{(0, 0, 0)\}$ and $(1, 1, 1) \rightsquigarrow (0, 0, 0)$, any non-constant EPIC mechanism is grossly Pareto inefficient, i.e. $\max_{s\in\Theta_+} q(s) \leq \min_{s\in\Theta_-} q(s)$.

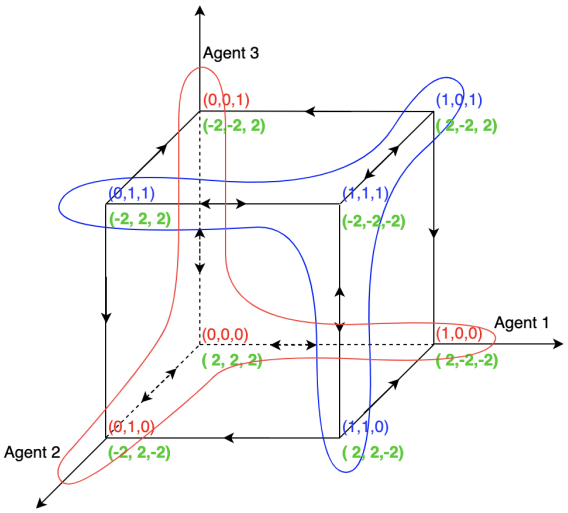
Hence, in this extreme example, any EPIC mechanism "cannot" response to any private information in the sense that responding to private information leads to efficiency loss.

A key feature of the environment is that though each agent i has an interpretation of each signal θ_j which is captured by the linear order $\succ_j^{(i)}$, agents disagree with each other regarding the interpretations of signals, for example, agent 1 thinks $\theta_1 = 1$ is a "good" signal for R, while agents 2 and 3 think $\theta_1 = 1$ is a "bad" signal for R. This observation leads to the following strengthening of monotonicity.

Definition I.26. *$\{u_i\}_{i\in I}$ is uniformly monotone if it is monotone and the linear order $\{\succ_j^{(i)}\}_{j\in I}$ is independent of i. That is, there is a collection of linear orders $\{\succ_j\}_{j\in I}$ on $\{\Theta_j\}_{j\in I}$ such that such that if $\theta_j \succ_i \theta'_j$, then*

$$sgn(u_i((\theta_i, \theta_{-i})) \geq sgn(u_i((\theta'_i, \theta_{-i}))$$

Figure I.3. An Example



for all θ_{-j} *and all* i.

Monotonicity requires that each agent i has clear (and her own) interpretation of θ_j. On top of that, uniform monotonicity makes sure every agent's interpretation is the same which is much stronger than what monotonicity asks for. However, both cases are common in reality. For example, most college admission officers would agree higher SAT scores could be the signal of competitive applicants; while different voters may have different tastes over political candidates' ideologies and policies.

Proposition I.15. *Suppose* $\{u_i(s)\}_{i\in I}$ *is uniformly monotone, then any EPIC mechanism is Pareto efficient in the range.*

I.7.2 Uniqueness

We focus on the existence of non-constant EPIC mechanisms in the main text. Here we provide a uniqueness result.

Definition I.27. *A mechanism* q *is a monotone transformation of* q' *if there is a weakly increasing function* $f : [0,1] \to [0,1]$ *such that*

$$q(\theta) = f(q'(\theta))$$

for all $\theta \in \Theta$.

Proposition I.16. *Suppose agents have common interests and $\{u_i\}_{i\in I}$ is uniformly monotone, then any non-constant EPIC mechanism q is a monotone transformation of q^*.*

Proposition I.16 states that, under uniform monotonicity and common interests, there is a unique class of non-constant EPIC mechanisms, and the Pareto efficient mechanism is one of them.

I.7.3 Optimal Design

Suppose there is a mechanism designer whose preference over mechanisms is represented by the utility function $d(q)$ that is linear in q:

$$d(q) = \sum_{s\in S} \delta(\theta) q(\theta).$$

Note that $\delta(\theta)$ can reflect the designer's own preference over R vs S in θ, or it can reflect the designer's concern for the agents' welfare. For instance, if $\delta(\theta) = \sum_{i=1}^{N} u_i(\theta)$ then d is utilitarian.

Fix a payoff environment, we denote EP the set of all EPIC mechanisms.

Definition I.28. *A mechanism $\bar{q}$ is optimal among all EPIC mechanisms if*

$$d(\bar{q}) = \sup_{q\in EP} d(q)$$

Lemma I.7. *There is an optimal EPIC mechanism that is deterministic.*

Lemma I.7 and Proposition I.2 jointly imply that optimal design is simplified to assigning $q_c \in \{0,1\}$ to every $c \in C^*$ respecting that $q_c \geq q_{c'}$ if $c \rightsquigarrow c'$.

Below we explicitly describe a procedure that returns an optimal mechanism.

Step 1. Construct the set $\mathcal{C}^B$ of all acyclic bipartitions of C^*.[15] Thus, any $\gamma \in C^*$ takes the form of $\gamma = \{C_1(\gamma), C_2(\gamma)\}$ where $c' \not\to c$ for any $c \in C_1(\gamma)$ and $c' \in C_2(\gamma)$.

Step 2. For every $\gamma \in \mathcal{C}^B$:

- If there exist $c \in C_1(\gamma)$ and $c' \in C_2(\gamma)$ such that $c \to c'$, then define $\hat{c}(\gamma) = \cup_{c\in C_1(\gamma)} c$.

[15] Lemma I.11 implies that $(C^*, \rightsquigarrow)$ corresponds to a directed acyclic graph (DAG). The construction of $\mathcal{C}^B$ is equivalent to finding all acyclic bipartitions of the DAG induced by $(C^*, \to)$.

– Otherwise, define $\hat{c}(\gamma) = \text{argmax}_{C \in \{C_1(\gamma), C_2(\gamma)\}} \sum_{c \in C} \sum_{s \in c} \delta(\theta)$.

Let $v(\gamma) := \sum_{s \in \hat{c}(\gamma)} \delta(\theta)$.

Step 3. Pick any $\gamma^* \in \text{argmax}_{\gamma \in \mathcal{C}^B} v(\gamma)$.

Step 4. Compare $v(\gamma^*)$ and $\max\{0, \sum_{\theta \in \Theta} \delta(\theta)\}$:

– If $v(\gamma^*) > \max\{0, \sum_{\theta \in \Theta} \delta(\theta)\}$, then return $\bar{q}$ where $\bar{q}(\theta) = 1$ if $\theta \in \hat{c}(\gamma^*)$ and $\bar{q}(\theta) = 0$ if $\theta \notin \hat{c}(\gamma^*)$.

– If $v(\gamma^*) \leq \max\{0, \sum_{\theta \in \Theta} \delta(\theta)\}$ and $\max\{0, \sum_{\theta \in \Theta} \delta(\theta)\} \geq 0$, then return $\bar{q}$ where $\bar{q}(\theta) = 1$ for every $\theta \in \Theta$.

– If $v(\gamma^*) \leq \max\{0, \sum_{\theta \in \Theta} \delta(\theta)\}$ and $\max\{0, \sum_{\theta \in \Theta} \delta(\theta)\} < 0$, then return $\bar{q}$ where $\bar{q}(\theta) = 0$ for every $\theta \in \Theta$.

Proposition I.17. *$\bar{q}$ is an optimal EPIC mechanism.*

I.8 Proof

This section collects proofs. We omit some proofs which are similar to others or straightforward.

I.8.1 DSIC

Proof of lemma I.1

Proof. "If": Suppose conditions 1 and 2 hold. Pick any $s \in \Theta_p$ and i. If $u_i(s) > 0$, then $\phi_i(\theta_i) \ni 1$, and hence for any θ'_{-i} and θ'_i:

$$u_i(s) q(\theta_i, \theta'_{-i}) = u_i(s) \max_{\theta''_i \in \Theta_i} q(\theta''_i, \theta'_{-i}) \geq u_i(s) q(\theta'_i, \theta'_{-i}).$$

The same equality for $u_i(s) \leq 0$ holds analogously. Therefore q is DSIC.

"Only if": Assume q is DSIC. Pick any i and $\theta_i \in \Theta_i$ where $\phi_i(\theta_i) \ni 1$. There is some $\theta_{-i} \in \Theta_{-i}$ where $(\theta_i, \theta_{-i}) \in \Theta$. Therefore $e_i(\theta_i, \theta_{-i}) q(\theta_i, \theta'_{-i}) \geq e_i(\theta_i, \theta_{-i}) q(\theta'_i, \theta'_{-i})$ for any θ'_i and θ'_{-i}, which implies that $q(\theta_i, \theta'_{-i}) \geq q(\theta'_i, \theta'_{-i})$ for any θ'_i and θ'_{-i}. Observe that condition 1 is satisfied in this case. Similarly, if $\phi_i(\theta_i) \ni -1$, then we have condition 2. □

Proof of Lemma I.2

Proof. Suppose q is DSIC and agent i is not privately informed. There exists θ_i such that $\{1,-1\} \subset \phi_i(\theta_i)$. From Lemma I.1 we have $\max_{\theta'_i \in \Theta_i} q(\theta'_i, \theta_{-i}) = q(\theta_i, \theta_{-i}) = \min_{\theta'_i \in \Theta_i} q(\theta'_i, \theta_{-i})$ for any θ_{-i}, which implies that $q(\cdot, \theta_{-i})$ is constant over Θ_i. □

Proof of Proposition I.1

Proof. Let $\Theta^{PI} := \times_{i \in PI} \Theta_i$. Therefore q is DSIC only if there is some $\hat{q} : \Theta^{PI} \to [0,1]$ such that $q(\theta) = \hat{q}(k(\theta))$ where $k : \Theta \to \Theta^{PI}$ orthogonally projects θ from Θ to Θ^{PI}.

We can then derive binary relation $\Rightarrow$ on S^{PI}: $\theta \Rightarrow \theta'$ if $\theta_i \Rightarrow_i \theta'_i$ for every $i \in PI$. Note that $\Rightarrow$ is a partial weak order.

"If": Suppose $q(\theta) = \hat{q}(k(\theta))$ for some $\hat{q} : \Theta^{PI} \to [0,1]$ where $\hat{q}(\theta) \geq \hat{q}(\theta')$ if $\theta \Rightarrow \theta'$. Pick any $\theta \in \Theta$, agent i, and $\theta'_i \in \Theta_i$ and $\theta'_{-i} \in \Theta_{-i}$. If $i \notin PI$ then $q(\theta_i, \theta'_{-i}) = \hat{q}(k(\theta_i, \theta'_{-i})) = \hat{q}(k(\theta'_i, \theta'_{-i})) = q(\theta'_i, \theta'_{-i})$, implying that i has no incentive to misreport θ'_i. Now consider $i \in PI$. If $\phi_i(\theta_i) = \{0\}$ then i is indifferent between S and R and hence has no incentive to misreport θ'_i. If $\phi_i(\theta_i) \ni 1$ then $u_i(\theta) \geq 0$ and also $k(\theta_i, \theta'_{-i}) \Rightarrow k(\theta'_i, \theta'_{-i})$. Therefore $u_i(\theta) q(\theta_i, \theta'_{-i}) = u_i(\theta) \hat{q}(k(\theta_i, \theta'_{-i})) \geq u_i(\theta) \hat{q}(k(\theta'_i, \theta'_{-i})) = u_i(\theta) q(\theta'_i, \theta'_{-i})$, that is, there is no incentive to misreport θ'_i. The same can be established analogously if $\phi_i(\theta_i) \ni -1$. Hence q is DSIC.

"Only if": Suppose q is DSIC. Lemma I.2 implies there is some $\hat{q} : \Theta^{PI} \to [0,1]$ such that $q(\theta) = \hat{q}(k(\theta))$. Fix any $\theta_{PI}, \theta'_{PI} \in \Theta^{PI}$ such that $\theta_{PI} \Rightarrow \theta'_{PI}$. We want to show that $\hat{q}(\theta_{PI}) \geq \hat{q}(\theta'_{PI})$, which is equivalent to showing that $q(\theta) \geq q(\theta')$ for any $\theta \in k^{-1}(\theta_{PI})$ and $\theta' \in k^{-1}(\theta'_{PI})$. For such θ, θ' consider the sequence $(\theta^0, ..., \theta^N)$ where θ^k agrees with θ' in the first k entries and with θ in the last $N-k$ entries. Therefore $\theta^0 = \theta$, $\theta^N = \theta'$, and θ^k and θ^{k-1} differ only in the k th entry. If $k \notin PI$ then $q(\theta^k) = q(\theta^{k-1})$ by Lemma I.2. If $k \in PI$ then we have $\theta_k \Rightarrow_k \theta'_k$, which implies that either $\phi_k(\theta_k^{k-1}) = \phi_k(\theta_k) \ni 1$ or $\phi_k(\theta_k^k) = \phi^k(\theta'_k) \ni -1$. In both cases we have $q(\theta^{k-1}) = q(\theta_k^{k-1}, \theta_{-k}^k) \geq q(\theta_k^k, \theta_{-k}^k) = q(\theta^k)$ by Lemma I.1. It follows that $q(\theta) = q(\theta^0) \geq q(\theta^1) \geq ... \geq q(\theta^N) = q(\theta')$. □

I.8.2 EPIC

Some Useful Lemmas

Lemma I.8. *q is EPIC if and only if $q(\theta) \geq q(\theta')$ for any θ, θ' where $\theta \to \theta'$.*

Proof of Lemma I.8

Proof. "If": Suppose $\theta \to \theta'$ implies $q(\theta) \geq q(\theta')$. Pick any state θ where $p(\theta) > 0$ and any agent i. If $u_i(\theta) = 0$, then i is indifferent between R and L, and hence she has no incentive to misreport in θ under any mechanism. If $u_i(\theta) > 0$, then we have $\theta \to (\theta'_i, \theta_{-i})$ for any $\theta'_i \neq \theta_i$. It follows that $q(\theta) \geq q(\theta'_i, \theta_{-i})$ for any $\theta'_i \in \Theta_i$, which implies that it is not profitable for i to misreport in θ, because i seeks to maximize the probability R being chosen in θ. Similarly it is not profitable for i to misreport if $u_i(\theta) < 0$. Therefore q is EPIC.

"Only if": Suppose q is EPIC. Pick any θ, θ' where $\theta \to \theta'$. By definition, there exists $i \in \{1, ..., N\}$ such that: (1) $\theta_{-i} = \theta'_{-i}$ and (2) $u_i(\theta) > 0$ or $u_i(\theta') < 0$. If $u_i(\theta) > 0$, then i prefers a higher probability of R being chosen in θ, and hence EPIC implies $\theta_i \in \text{argmax}_{\hat{\theta}_i \in \Theta_i} q(\hat{\theta}_i, \theta_{-i})$, which in turn implies that $q(\theta) = q(\theta_i, \theta_{-i}) \geq q(\theta'_i, \theta_{-i}) = q(\theta')$. The case where $u_i(\theta') < 0$ is analogous. □

Lemma I.9 is a corollary of Lemma I.8.

Lemma I.9. *If q is EPIC if and only if $q(\theta) \geq q(\theta')$ for any θ, θ' where $\theta \rightsquigarrow \theta'$.*

The following lemma will be useful.

Lemma I.10. *For any partition C of S, $c, c' \in C$, $\theta \in c$ and $\theta' \in c'$, $c \rightsquigarrow c'$ if $\theta \rightsquigarrow \theta'$. Moreover, if $C = C^*$ then $c \rightsquigarrow c'$ only if $\theta \rightsquigarrow \theta'$*

Proof. Suppose $\theta \rightsquigarrow \theta'$, then there exists a list $(\theta^0, ..., \theta^J)$ such that $\theta^0 = \theta \to \theta^1 ... \to \theta^J = \theta'$. Let c^j denote the cell in C that contains θ^j, and hence $c = c^0 \to c^1 ... \to c^J = c'$, which implies that $c \rightsquigarrow c'$.

Suppose $C = C^*$ and $c \to c'$. By definition there exist $\hat{\theta} \in c$ and $\hat{\theta}' \in c'$ such that $\hat{\theta} \to \hat{\theta}'$, which implies that $\hat{\theta} \rightsquigarrow \hat{\theta}'$. By construction of C^*, $\theta \rightsquigarrow \hat{\theta}$ and $\hat{\theta}' \rightsquigarrow \theta'$, hence $\theta \rightsquigarrow \theta'$ because $\rightsquigarrow$ is transitive. With a straightforward inductive argument it is easy to generalize this observation as long as $c \rightsquigarrow c'$. □

We say that a partition C of S is **acyclic** if there does not exist distinct $c, c' \in S$ such that $c \leftrightsquigarrow c'$.

Lemma I.11. *C^* is the finest acyclic partition of Θ.*

Proof. For any $c, c' \in C^*$, if $c \leftrightsquigarrow c'$ then $\theta \leftrightsquigarrow \theta'$ for any $\theta \in c$ and $\theta' \in c'$ by Lemma I.10, which implies that $c = c'$ by the construction of C^*. Thus there does not exist distinct $c, c' \in \Theta$ such that $c \leftrightsquigarrow c'$, implying that C^* is acyclic.

Now we show that C^* is the finest acyclic partition of Θ. Consider another acyclic partition $\overline{C}$ of Θ. Pick any $c \in C^*$. Suppose, in order to lead to a contradiction, that there

are distinct $\overline{c}, \overline{c}' \in \overline{C}$ both of which intersect with c. Pick $\theta \in c \cap \overline{c}$ and $\theta' \in c \cap \overline{c}'$. It follows that $\theta \leftrightsquigarrow \theta'$ by construction of C^*, and hence $\overline{c} \leftrightsquigarrow \overline{c}'$ by Lemma I.10, a contradiction as $\overline{C}$ is assumed to be acyclic. Therefore $c \subset \overline{c}$ for some $\overline{c} \in \overline{C}$, implying that C^* is finer than C. □

Proof of Proposition I.2

Proof. "If": If there are such probabilities then $\theta \rightsquigarrow \theta'$ implies $C^*(\theta) \rightsquigarrow C^*(\theta')$ (where $C^*(\theta)$ denotes the cell in C^* that contains θ) and hence $q(\theta) = q_{C^*(\theta)} \geq q_{C^*(\theta')} = q(\theta')$, which by Lemma I.9 implies q is EPIC.

"Only if": If q is EPIC then Lemma I.9 implies that $q(\theta) = q(\theta')$ if $C^*(\theta) = C^*(\theta')$, because $\theta \rightsquigarrow \theta' \rightsquigarrow \theta$. Therefore q is constant on any $c \in C^*$, and we can denote this value as q_c. Moreover if $c \rightsquigarrow c'$ for $c, c' \in C^*$ then Lemma I.10 implies that $\theta \rightsquigarrow \theta'$ for any $\theta \in c$ and $\theta' \in c'$ and hence $q(\theta) \geq q(\theta')$ by Lemma I.9, implying $q_c \geq q_{c'}$. □

Proof of Proposition I.3

Proof. That part 1 and part 2 are equivalent follows immediately from Lemma I.2.

Suppose part 2 is true. That C^* is acyclic implies that there exists $c \in C^*$ such that $c' \not\rightsquigarrow c$ for every $c' \in C^*$. Let $\hat{c}$ denote the set in C^* containing c. Construct $c_A = \{c' : c' \in C \text{ and } c' \rightsquigarrow \hat{c}\}$ and $c_B := \Theta \backslash c_A$. By construction $c_B \not\rightsquigarrow c_A$, proving part 3.

That part 3 implies part 4 is obvious.

Suppose part 4 is true. There exists $\theta \in \Theta$ such that $\theta' \not\rightsquigarrow \theta$ for some $\theta' \in \Theta_p$. Construct $c_A := \{\theta' : \theta' \in \Theta \text{ and } \theta' \rightsquigarrow \theta\}$, and $c_B := \Theta \backslash c_A$. Note that by construction $c_B \not\rightsquigarrow c_A$, hence $C = \{c_A, c_B\}$ is an acyclic partition of S where both elements intersect with Θ_p. Since C^* is a refinement of C by Lemma I.11, part 2 follows. □

Proof of Proposition I.15

Definition I.29. *A vertex θ is called a source if there is a walk from θ to θ' for any $\theta' \in \Theta$; A vertex θ is called a sink if there is a walk from θ' to θ for any $\theta' \in \Theta$.*

When $\{u_i\}_{i \in I}$ is weakly monotone, we can find $\bar{\theta} = (\bar{\theta}_1, \bar{\theta}_2, \ldots, \bar{\theta}_N)$ such that $\bar{\theta}_i \succ_i^{(i)} \theta_i$ for all $\theta_i \neq \bar{\theta}_i$ and $\underline{\theta} = (\underline{\theta}_1, \ldots, \underline{\theta}_N)$ such that $\theta_i \succ_i^{(i)} \underline{\theta}_i$ for all $\theta_i \neq \bar{\theta}_i$. Alternatively, we can rearrange Θ_i by $\succ_i^{(i)}$ such that $\theta_i^{(1)} \succ_i^{(i)} \theta_i^{(2)} \succ_i^{(i)} \cdots$, then $\bar{\theta} = (\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_N^{(1)})$.

Lemma I.12. *Suppose $\{u_i(\theta)\}_{i \in I}$ is monotone, then $\bar{\theta}$ is a source and $\underline{\theta}$ is a sink.*

Lemma I.13. *Suppose $\{u_i(\theta)\}_{i\in I}$ is uniformly monotone, then*

1. *if there is θ such that $u_i(\theta) > 0$ for all $i \in I$ then there is a $\bar{\theta} - \theta$ walk;*

2. *if there is θ such that $u_i(\theta) < 0$ for all $i \in I$ then there is a $\theta - \underline{\theta}$ walk;*

Proof of Proposition I.4

Proof. We want to show that for any $\theta \in \Theta$, $i \in I$ and $\theta_i' \in \Theta_i$,

$$u(\theta)q(\theta_i, \theta_{-i}) \geq u(\theta)q(\theta_i', \theta_{-i}).$$

Suppose $u(\theta) > 0$, then above equation becomes $q(\theta) = 1 \geq q(\theta_i', \theta_{-i})$. Suppose $u(\theta) < 0$, then above equation becomes $q(\theta) = 0 \leq q(\theta_i', \theta_{-i})$. Suppose $u(\theta) = 0$, then above equation becomes $0 \geq 0$. □

Proof of Lemma I.7

Proof. Given Lemma I.8, it is straightforward to verify that the set of EPIC mechanisms is a polytope in $\mathbb{R}^{|S|}$ with extreme points consisting of either 0s or 1s. Since the objective function d is linear in q, there is a maximizer that coincides with an extreme point, which is then a deterministic mechanism. □

Proof of Proposition I.17

Proof. For any $\gamma \in \mathcal{C}^B$, define $q_\gamma : \Theta \to \{0, 1\}$ such that $q_\gamma(\theta) = 1$ if $\theta \in \hat{c}(\gamma)$ and $q_\gamma(\theta) = 0$ if $\theta \notin \hat{c}(\gamma)$.

Lemma I.2 imply that any non-constant deterministic EPIC mechanism must be measurable with respect to some $\gamma \in \mathcal{C}^B$, i.e. $q^{-1}(x) \subset \gamma$ for $x = 0, 1$. Moreover, if there exist $c \in C_1(\gamma)$ and $c' \in C_2(\gamma)$ such that $c \to c'$, then Lemma I.2 implies that q_γ is the only non-constant deterministic EPIC mechanism measurable with respect to γ. Otherwise, there are two non-constant deterministic EPIC mechanisms measurable with respect to γ: q_γ^1 that assigns 1 to states in every $c \in C_1(\gamma)$ and 0 to states in every $c \in C_2(\gamma)$, and q_γ^2 that assigns 0 to states in every $c \in C_1(\gamma)$ and 1 to states in every $c \in C_2(\gamma)$. Observe that, in this case q_γ is set to be the mechanism that the designer prefers between q_γ^1 or q_γ^2.

It follows that every optimal non-constant deterministic EPIC mechanism is in $\operatorname{argmax}_{\gamma\in\mathcal{C}^B} d(q_\gamma)$. It is straightforward to verify that $v(\gamma) = d(q_\gamma)$. Thus γ^* is an optimal non-constant deterministic EPIC mechanism.

Step 4 simply compares an optimal non-constant deterministic EPIC mechanism to a constant deterministic mechanism (which must be EPIC), and yields the optimal deterministic EPIC mechanism as $\bar{q}$. It then follows from Lemma I.7 that $\bar{q}$ is an optimal EPIC mechanism. □

Proof of Proposition I.5

Proof. Suppose a fully reduced mechanism q is EPIC. We want to show q is constant over $\bar{\Theta} := \Theta \setminus (IC_1 \cup IC_2)$ where $\bar{\Theta}$ is the set of payoff states in which both agents have strict ex post preferences.

We prove the Proposition in four steps.

Step 1: We introduce a binary relation $\sim$ on $\bar{\Theta}$.

Definition I.30. *We Two payoff states $\theta \sim \theta'$ if*

1. *$u_i(\theta)u_i(\theta') > 0$ for $i = 1, 2$;*
2. *There exists a path, $(\theta = \theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(n)} = \theta')$, such that $\theta^{(k)}$ and $\theta^{(k+1)}$ differ in one entry, and $\theta'' \in \{\theta | \theta = t\theta^{(k)} + (1-t)(\theta^{(k+1)}), \forall k = 1, 2, ..., n, \forall t \in [0, 1]\}$ implies $u_i(\theta'')u_i(\theta'') > 0$ for $i = 1, 2$.*

Condition 1 means each agent has the same ex post preferences on θ and θ'. Condition 2 means thee exist a continuous "manhattan path" links θ and θ' and all the payoff states along the path give each agent the same ex post preferences as θ and θ' give. It is easy to verify that $\sim$ is reflexive, symmetric, and transitive. Hence, $\sim$ is an equivalence relation which induces in partition $P := \{P_k\}$ on $\bar{\Theta}$. If P is a singleton, then Proposition I.5 holds trivially. From now on, we assume P contains at least two elements/blocks.

Step 2: We show that q is constant in each P_i.

For any $\theta, \theta' \in P_i$, there exists a path $(\theta = \theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(n)} = \theta')$ links θ and θ'. Ex post incentive compatibility requires $q(\theta^{(k)}) = q(\theta^{(k+1)})$. Hence, $q(\theta) = q(\theta')$.

Step 3: We prove that q is constant between adjacent P_i and P_j.

We first give the definition of adjacency. Note that u_1 and u_2 are continuous, each P_i is an open set in Θ. We denote the closure of P_i by $\bar{P}_i$.

Definition I.31. *Two blocks of the partition P, P_i, P_j, are adjacent if $m(\bar{P}_i \cap \bar{P}_j) > 0$ where m is the Lebesgue measure of $\mathbb{R}$.*

That is, $P_i, P_j \in P$ are adjacent if their closures intersect with each other in a non-degenerated fashion. Hence, we can find $\theta \in \bar{P}_i \cap \bar{P}_j$ and $\delta > 0$ such that $B_\delta(\theta) \subset \bar{P}_i \cup \bar{P}_j$ where $B_\delta(\theta)$ is the $\delta-$neighborhood of θ under the Euclidean norm. We discuss three cases.

Case 1: P_i and P_j share the same signs of (u_1, u_2).

Since $m(\bar{P}_i \cap \bar{P}_j) > 0$, there exist $\theta' \in P_i$ and $\theta'' \in P_j$ such that $\text{sgn}u_1(\theta') = \text{sgn}u_1(\theta'')$, $\text{sgn}u_2(\theta') = \text{sgn}u_2(\theta'')$, and θ' and θ'' differ in one entry. Without loss of generality, we assume $\theta'_1 = \theta''_1$. Then, by agent 1's incentive compatibilities, $q(\theta') = q(\theta'')$.

Case 2: P_i and P_j differ in one sign of (u_1, u_2).

Without loss of generality, we assume they differ in u_1. Then $\bar{P}_i \cap \bar{P}_j$ is a subset of $BD_1 \setminus BD_2$. For the θ we found right after the Definition I.31, we know $\theta \in \bar{P}_i \cap \bar{P}_j \subset BD_1$, we have $\dfrac{\partial u_1(\theta)}{\partial \theta_2} \neq 0$ by generic interdependence. Therefore, $u_1(\theta_1, \theta_2 + \delta/2)u_1(\theta_1, \theta_2 - \delta/2) < 0$. Then we know $u_2(\theta_1, \theta_2+\delta/2)u_2(\theta_1, \theta_2-\delta/2) > 0$. Agent 2's incentive compatibilities require that $q(\theta_1, \theta_2+\delta/2) = q(\theta_1, \theta_2-\delta/2)$. Since $(\theta_1, \theta_2+\delta/2)$ and $(\theta_1, \theta_2-\delta/2)$ belongs to different blocks, q is constant between P_i and P_j follows immediately.

Case 3: P_i and P_j differ in both signs of (u_1, u_2).

Suppose $u_1(\theta') > 0$, $u_2(\theta') > 0$ for all $\theta' \in P_i$. Then $u_1(\theta'') < 0$, $u_2(\theta'') < 0$ for all $\theta'' \in P_j$. Thus $\theta \in (\bar{P}_i \cap \bar{P}_j) \subset (BD_1 \cap BD_2)$. It is easy to show that all for all $\theta' \in B_\delta(\theta)$, we have $u_1(\theta')u_2(\theta') > 0$. Generic heterogeneity condition is violated. Similarly, the subcase $u_1(\theta') < 0$, $u_2(\theta') < 0$ violates generic heterogeneity condition.

Now suppose $u_1(\theta') < 0$, $u_2(\theta') > 0$ for all $\theta' \in P_i$. Then $u_1(\theta'') > 0$, $u_2(\theta'') < 0$ for all $\theta'' \in P_j$. Then $(\bar{P}_i \cap \bar{P}_j) \subset (BD_1 \cap BD_2)$. We can pick two payoff states within the δ-neighborhood of θ: $(\theta_1 - \epsilon, \theta_2)$, $(\theta_1 + \epsilon, \theta_2)$. We know $u_1(\theta_1 - \epsilon, \theta_2)u_1(\theta_1 + \epsilon, \theta_2) < 0$. Without loss of generality, we assume $u_1(\theta_1 - \epsilon, \theta_2) < 0$ and $u_1(\theta_1 + \epsilon, \theta_2) > 0$. Then, $(\theta_1 - \epsilon, \theta_2) \in P_i$, $(\theta_1 + \epsilon, \theta_2) \in P_j$, and $q(\theta_1 - \epsilon, \theta_2) \leq q(\theta_1 + \epsilon, \theta_2)$.

We can also pick another two payoff states within the δ-neighborhood of $(\theta_1, \theta_2 - \epsilon')$, $(\theta_1, \theta_2 + \epsilon')$. Similarly, we know $u_2(\theta_1, \theta_2 - \epsilon')u_2(\theta_1, \theta_2 - \epsilon') < 0$. Without loss of generality, we assume $u_2(\theta_1, \theta_2 - \epsilon') > 0$ and $u_\epsilon(\theta_1, \theta_2 + \epsilon') < 0$. Then, $(\theta_1, \theta_2 - \epsilon') \in P_i$, $(\theta_1, \theta_2 + \epsilon') \in P_j$, and $q(\theta_1, \theta_2 - \epsilon') \geq q(\theta_1, \theta_2 + \epsilon')$. By step 2, we know $q(\theta_1 - \epsilon, \theta_2) = q(\theta_1, \theta_2 - \epsilon')$, $q(\theta_1 + \epsilon, \theta_2) = q(\theta_1, \theta_2 + \epsilon')$. Therefore, $q(\theta_1 - \epsilon, \theta_2) = q(\theta_1, \theta_2 - \epsilon') = q(\theta_1 + \epsilon, \theta_2) = q(\theta_1, \theta_2 + \epsilon')$.

The case $u_1(\theta') > 0$, $u_2(\theta') < 0$ for all $\theta' \in P_i$ can be proved in the same way.

Step 4: Since every block P_i has at least one adjacent block, any two blocks P_i and P_j are linked by a sequence of adjacent blocks. Hence, q is constant over $\bar{\Theta}$. □

I.8.3 IDSIC

Proof of Lemma I.4

Proof. For player i of type t_i, given message profile $m_{-i} \in M_{-i}$, payoff-type profile $\theta_{-i} \in \Theta_{-i}$ and joint strategy $\sigma_{-i} : T_{-i} \to \Delta(M_{-i})$ of the other players, define

$$q_i(m_{-i}|\sigma_{-i}, \theta_{-i}, t_i) := \sum_{\{t_{-i} \in T_{-i} : \hat{\theta}_{-i}(t_{-i}) = \theta_{-i}\}} \Pr\left[t_{-i} \middle| \theta_{-i}, t_i\right] \sigma_{-i}(t_{-i})[m_{-i}]$$

where $\Pr\left[t_{-i} \middle| \hat{\theta}_{-i}(t_{-i}) = \theta_{-i}, t_i\right] = \dfrac{\hat{\beta}_i(t_i)[t_{-i}]}{\hat{b}_i(t_i)[\theta_{-i}]}$ is agent i's belief over t_{-i} conditional on her own signal is t_i and other agents' payoff types are θ_{-i}.

Thus $q_i(m_{-i}|\sigma_{-i}, \theta_{-i}, t_i)$ is what player i of type t_i evaluates as the probability that the message profile from the other players will be m_{-i} conditional on their payoff-type profile being θ_{-i} and them following σ_{-i}.

For player i of type t_i, her expected payoff from message m_i conditional on other players following joint strategy σ_{-i} can be written as:

$$\begin{aligned} &U_i(m_i|\sigma_{-i}, t_i) \\ &\quad := \sum_{\theta_{-i} \in \Theta_{-i}} \hat{b}_i(t_i)[\theta_{-i}] \sum_{m_{-i} \in M_{-i}} q_i(m_{-i}|\sigma_{-i}, \theta_{-i}, t_i) \sum_{a \in A} u_i(a, \hat{\theta}_i(t_i), \theta_{-i}) q(m_i, m_{-i})[a]. \end{aligned}$$

Suppose m_i is an interim dominant action for player i of type t_i. Pick any $t_i' \in T_i$ where $\hat{\theta}_i(t_i') = \hat{\theta}_i(t_i)$ and $\hat{b}_i(t_i) = \hat{b}_i(t_i')$. For any $\sigma_{-i} : T_{-i} \to \Delta(M_{-i})$ there always exists $\chi_{\sigma_{-i}} : T_{-i} \to \Delta(M_{-i})$ such that $q_i(m_{-i}|\sigma_{-i}, \theta_{-i}, t_i') = q_i(m_{-i}|\chi_{\sigma_{-i}}, \theta_{-i}, t_i)$ for any $m_{-i} \in M_{-i}$ and $\theta_{-i} \in \Theta_{-i}$.[16]

It is straightforward to verify that $U_i(\cdot|\sigma_{-i}, t_i') = U_i(\cdot|\chi_{\sigma_{-i}}, t_i)$. It follows that for any σ_{-i} and m_i',

$$U_i(m_i|\sigma_{-i}, t_i') = U_i(m_i|\chi_{\sigma_{-i}}, t_i) \geq U_i(m_i'|\chi_{\sigma_{-i}}, t_i) = U_i(m_i'|\sigma_{-i}, t_i').$$

Thus m_i is also an interim dominant action for player i when her type is t_i'. □

Proof of Proposition I.6

Proof. **"Only if'."** Suppose σ is an interim dominant-strategy equilibrium of the mechanism. For each $i \in I$, $\theta_i \in \Theta_i$ and $b_i \in \Delta(\Theta_{-i})$ pick some $\tau_i(\theta_i, b_i) \in T_i$ where $\hat{\theta}_i(\tau_i(\theta_i, b_i)) = \theta_i$ and

[16] For instance, set $\chi_{\sigma_{-i}}(m_{-i}|t_{-i}) = q_i(m_{-i}|\sigma_{-i}, \theta_{-i}, t_i')$ for every t_{-i} where $\hat{\theta}_{-i}(t_{-i}) = \theta_{-i}$.

$\hat{b}_i(\tau_i(\theta_i, b_i)) = b_i$ whenever possible. Consider strategy profile σ' where $\sigma'_i(t_i) = \sigma'_i(\tau_i(\theta_i, b_i))$ for any t_i where $\hat{\theta}_i(t_i) = \theta_i$ and $\hat{b}_i(t_i) = b_i$. Observe that σ' is higher-order belief-independent. Moreover Lemma I.4 implies that σ' is also an interim dominant strategy equilibrium. □

Proof of Lemma I.6

Proof. **"If":** Fix agent i. Suppose the other agents report $\sigma_{-i}(\theta_{-i}) \in H_{-i}$ when their true payoff types is θ_{-i}. The difference in payoff to i between truthfully reporting h_i and misreporting as h'_i is

$$\begin{aligned} D :=& \sum_{\theta_{-i} \in \Theta_{-i}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}) \big(q(h_i, \sigma_{-i}(\theta_{-i})) - q(h'_i, \sigma_{-i}(\theta_{-i}))\big) \\ =& \Big[\sum_{\{\theta_{-i} | u_i(\theta) > 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}) \big(q(h_i, \sigma_{-i}(\theta_{-i})) - q(h'_i, \sigma_{-i}(\theta_{-i}))\big) \\ &+ \sum_{\{\theta_{-i} | u_i(\theta) < 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}) \big(q(h_i, \sigma_{-i}(\theta_{-i})) - q(h'_i, \sigma_{-i}(\theta_{-i}))\big) \Big]. \end{aligned}$$

Define $\overline{d} := \max_{h_{-i} \in H_{-i}} \big(q(h_i, h_{-i}) - q(h'_i, h_{-i})\big)$ and $\underline{d} := \min_{h_{-i} \in H_{-i}} \big(q(h_i, h_{-i}) - q(h'_i, h_{-i})\big)$. It is straightforward to verify that

$$\begin{aligned} D &\geq \overline{d} \sum_{\{\theta_{-i} | u_i < 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}) + \underline{d} \sum_{\{\theta_{-i} | u_i > 0\}} b_i(\theta_{-i}) u_i(\theta_i, \theta_{-i}) \\ &= \underline{\alpha}_i(h_i) \overline{d} + \overline{\alpha}_i(h_i) \underline{d} \\ &\geq 0, \end{aligned}$$

where the last inequality is due to the assumption that for any $h_{-i}, h'_{-i} \in H_{-i}$:

$$\underline{\alpha}_i(h_i) \Big(q(h_i, h_{-i}) - q(h'_i, h_{-i})\Big) + \overline{\alpha}_i(h_i) \Big(q(h_i, h'_{-i}) - q(h'_i, h'_{-i})\Big) \geq 0.$$

Therefore misreporting is not profitable, and q is IDSIC.

"Only if": Suppose q is IDSIC. Fix agent i, $h_i, h'_i \in H_i$ and $h_{-i}, h'_{-i} \in H_{-i}$. Suppose agents other than i jointly report h_{-i} whenever their true payoff types $\hat{\theta}_{-i}$ satisfy $u_i(\theta_i, \hat{\theta}_{-i}) < 0$, and they jointly report h'_{-i} otherwise. Thus IDSIC requires that

$$\begin{aligned} &\sum_{\{\hat{\theta}_{-i} | u_(\theta_i, \hat{\theta}_{-i}) < 0\}} b_i(\theta_{-i}) u_i(\theta_i, \hat{\theta}_{-i}) \big(q(h_i, h_{-i}) - q(h'_i, h_{-i})\big) \\ &\quad + \sum_{\{\hat{\theta}_{-i} | u_i(\theta_i, \hat{\theta}_{-i}) > 0\}} b_i(\theta_{-i}) u_i(\theta_i, \hat{\theta}_{-i}) \big(q(h_i, h'_{-i}) - q(h'_i, h'_{-i})\big) \geq 0, \end{aligned}$$

implying that $\underline{\alpha}_i(h_i)\big(q(h_i,h_{-i})-q(h_i',h_{-i})\big)+\overline{\alpha}_i(h_i)\big(q(h_i,h'_{-i})-q(h_i',h'_{-i})\big)\geq 0$. □

Proof of Proposition I.8

Proof. **"Only if":** Suppose q is IDSIC. Then by Lemma I.6, Equation I.2 holds for all $h_i, h_i', h_{-i}, h'_{-i}$. Let $h_{-i}=h'_{-i}$, we have

$$\begin{aligned}\underline{\alpha}_i(h_i)\big(q(h_i,h_{-i})-q(h_i',h_{-i})\big)\geq&-\overline{\alpha}_i(h_i)\big(q(h_i,h_{-i})-q(h_i',h_{-i})\big)\\ \Leftrightarrow\alpha_i(h_i)q(h_i,h_{-i})\geq&\alpha_i(h_i)q(h_i',h_{-i})\end{aligned}\tag{I.3}$$

for all $h_i, h_i' \in H_i$ and all $h_{-i}\in H_{-i}$.

If $\alpha_i(h_i)>0$, then $q(h_i,h_{-i})\geq q(h_i',h_{-i})$ for all $h_i'\in H_i$; if $\alpha_i(h_i)<0$, then $q(h_i,h_{-i})\leq q(h_i',h_{-i})$ for all $h_i'\in H_i$; if $\alpha_i(h_i)=0$, then $q(h_i',h_{-i})\leq q(h_i,h_{-i})\leq q(h_i'',h_{-i})$ for all $h_i'\in H_i^-, h_i'\in H_i^+$. Hence, the monotonicity condition is proved.

For each agent i, we have two cases for the variation condition (condition 2).

1. $H_i^0=\emptyset$. First, let $h_i\in H_i^+, h_i'\in H_i^-$, $h_{-i}\in \operatorname{argmax}[q(h_i,h_{-i})-q(h_i',h_{-i})]$, $h'_{-i}\in \operatorname{argmin}[q(h_i,h_{-i})-q(h_i',h_{-i})]$. Equation I.2 gives

$$-\underline{\alpha}_i(h_i)\max_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)\leq\overline{\alpha}_i(h_i)\min_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big),$$

for all $h_i\in H_i^+$. Therefore,

$$\max_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)\leq\rho_i(h_i)\min_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big),$$

for all $h_i\in H_i^+$.

Second, let $h_i\in H_i^-, h_i'\in H_i^+, h_{-i}\in\operatorname{argmin}[q(h_i,h_{-i})-q(h_i',h_{-i})]$, $h'_{-i}\in\operatorname{argmax}[q(h_i,h_{-i})-q(h_i',h_{-i})]$. Equation I.2 gives

$$-\underline{\alpha}_i(h_i)\min_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)\geq\overline{\alpha}_i(h_i)\max_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)$$

Therefore,

$$\max_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)\leq\rho_i(h_i)\min_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big),$$

for all $h_i\in H_i^-$. Recall that $\rho_i=\min_{h_i}\rho_i(h_i)$. Hence,

$$\max_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big)\leq\rho_i\min_{h_{-i}\in H_{-i}}\Big(\overline{q}_i(h_{-i})-\underline{q}_i(h_{-i})\Big).$$

2. $H_i^0 \neq \emptyset$. Now consider $h_i^* \in H_i^0$. Since $-\overline{\alpha}_i(h_i^*) = \overline{\alpha}_i(h_i^*)$, Equation I.2 gives

$$q(h_i^*, h_{-i}) - q(h_i', h_{-i}) \leq q(h_i^*, h_{-i}') - q(h_i', h{-}i')$$

for all $h_i' \in H_i$, and all $h_{-i}, h_{-i}' \in H_{-i}$. Then $q(h_i^*, h_{-i}) - q(h_i', h_{-i}) = q(h_i^*, h_{-i}') - q(h_i', h_{-i}')$ for all $h_i' \in H_i$ and all $h_{-i}, h_{-i}' \in H_{-i}$, which mean $q(h_i^*, h_{-i}) - q(h_i', h_{-i})$ is constant over h_{-i}, for all $h_i' \in H_i$. We then have $q(h_i, h_{-i}) - q(h_i', h_{-i}) = [q(h_i, h_{-i}) - q(h_i^*, h_{-i})] + [q(h_i^*, h_{-i}) - q(h_i', h_{-i})]$ is a constant function with respect to h_{-i}, for all $h_i, h_i' \in H_i$.

"If": We want to show that if both monotonicity condition and variation condition are satisfied, then Equation I.2 holds for all $i \in I$, $h_i, h_i' \in H_i$, and $h_{-i}, h_{-i}' \in H_{-i}$. For each $i \in I$, we discuss two situations.

1. $H_i^0 = \emptyset$. If both $h_i, h_i' \in H_i^+$ or both $h_i, h_i' \in H_i^-$, then it is obvious that Equation I.2 holds. If $h_i \in H_i^+, h_i' \in H_i^-$, then

$$\begin{aligned}
&\underline{\alpha}_i(h_i)\big(q(h_i, h_{-i}) - q(h_i', h_{-i})\big) + \overline{\alpha}_i(h_i)\big(q(h_i, h_{-i}') - q(h_i', h_{-i}')\big) \\
=&\,\underline{\alpha}_i(h_i)\big(\overline{q}(h_{-i}) - \underline{q}(h_{-i})\big) + \overline{\alpha}_i(h_i)\big(\overline{q}(h_{-i}') - \underline{q}(h_{-i}')\big) \\
\geq&\,\underline{\alpha}_i(h_i) \max_{h_{-i}} \big(\overline{q}(h_{-i}) - \underline{q}(h_{-i})\big) + \overline{\alpha}_i(h_i) \min_{h_{-i}'} \big(\overline{q}(h_{-i}') - \underline{q}(h_{-i}')\big) \\
\geq&\, 0
\end{aligned}$$

for all $h_{-i}, h_{-i}' \in H_{-i}$. The first equality follows the monotonicity condition, the second equality is due to the fact $\underline{\alpha}_i(h_i) \leq 0$ and $\overline{\alpha}_i(h_i) > 0$, the last inequality follows the variation condition. The case in which $h_i \in H_i^-, h_i' \in H_i^+$ can be prove in a similar way.

2. $H_i^0 \neq \emptyset$. Then variation condition says $q(h_i, h_{-i}) - q(h_i', h_{-i})$ is a constant function with respect to h_{-i}, for all $h_i, h_i' \in H_i$. Equation I.2 becomes

$$\alpha_i(h_i)[q(h_i, h_{-i}) - q(h_i', h_{-i})] \geq 0$$

which is true for all $h_i, h_i' \in H_i$ and $h_{-i} \in H_{-i}$ by the monotonicity condition.

□

Proof of Proposition I.9

Proof. **"If":** Fix any type space T and additive mechanism $(M_1, ..., M_N, q)$. It suffices to show that every agent has an interim dominant strategy. For agent i, consider the strategy

σ_i that prescribes her to send a message $\overline{m}_i$ that maximizes $\pi_i^q(\cdot)$ if her type t_i satisfies $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) \geq 0$, or to send a message $\underline{m}_i$ that minimizes $\pi_i^q(\cdot)$ if her type satisfies $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) < 0$. For any strategy profile σ_{-i} from the other agents, if $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) \geq 0$, then we have

$$
\begin{aligned}
U_i(\overline{m}_i|\sigma_{-i}, t_i) &\geq U_i(m_i|\sigma_{-i}, t_i) \\
\Leftrightarrow \sum_{t_{-i}\in T_i} \hat{\beta}_i(t_i)[t_{-i}]q(\overline{m}_i, \sigma_{-i}(t_{-i}))u_i(\hat{\theta}(t_i, t_{-i})) &\geq \sum_{t_{-i}\in T_i} \hat{\beta}_i(t_i)[t_{-i}]q(m_i, \sigma_{-i}(t_{-i}))u_i(\hat{\theta}(t_i, t_{-i})) \\
\Leftrightarrow \sum_{t_{-i}\in T_i} \hat{\beta}_i(t_i)[t_{-i}]\pi_i^q(\overline{m}_i)u_i(\hat{\theta}(t_i, t_{-i})) &\geq \sum_{t_{-i}\in T_i} \hat{\beta}_i(t_i)[t_{-i}]\pi_i^q(m_i)u_i(\hat{\theta}(t_i, t_{-i})) \\
\Leftrightarrow \pi_i^q(\overline{m}_i)\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) &\geq \pi_i^q(m_i)\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) \\
\Leftarrow \pi_i^q(\overline{m}_i) &\geq \pi_i^q(m_i),
\end{aligned}
$$

from which we conclude that $\overline{m}_i$ is a best response against σ_{-i}. That $\underline{m}_i$ is a best response against σ_{-i} when $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) < 0$ is established analogously. Therefore σ_i is indeed an interim dominant strategy for agent i.

"Only if": Suppose $(M_1, ..., M_N, q)$ is IDSIC in all type spaces. We want to show that there exist functions $\pi_i^q : M_i \to [0, 1]$ such that $q(m_1, ..., m_N) = \sum_{i\in I} \pi_i^q(m_i)$.

Fix a type space where for every $i \in I$ there is a type $\tilde{t}_i$ such that $\overline{\alpha}_i(\hat{\theta}_i(\tilde{t}_i), \hat{b}_i(\tilde{t}_i)) = -\underline{\alpha}_i(\hat{\theta}_i(\tilde{t}_i), \hat{b}_i(\tilde{t}_i)) > 0$. Let m_i^* denote the message that agent i of type $\tilde{t}_i$ sends in a given interim dominant strategy equilibrium. It follows from an argument analogous to how Lemma I.6 is proved that

$$
\underline{\alpha}_i(\hat{\theta}_i(\tilde{t}_i), \hat{b}_i(\tilde{t}_i))\big(q(m_i^*, m_{-i}) - q(m_i, m_{-i})\big) + \overline{\alpha}_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i))\big(q(m_i^*, m'_{-i}) - q(m_i, m'_{-i})\big) \geq 0
$$

for any $m_i \in M_i, m_{-i}, m'_{-i} \in M_{-i}$. That $\underline{\alpha}_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) + \overline{\alpha}_i(\hat{\theta}_i(\tilde{t}_i), \hat{b}_i(\tilde{t}_i)) = 0$ implies $q(m_i^*, m_{-i}) - q(m_i, m_{-i})$ in invariant to m_{-i}. Thus the expression

$$
q(m_i, m_{-i}) - q(m_i^*, m_{-i}) + \frac{1}{N} q(m_1^*, ..., m_N^*).
$$

does not depend on m_{-i}, and hence we can denote this expression as $\pi_i^q(m_i)$. Observe that

for any $m_1, ..., m_N$,

$$\begin{aligned} q(m_1, ..., m_N) = \; & q(m_1, m_2, ..., m_N) - q(m_1^*, m_2, ..., m_N) + \frac{1}{N} q(m_1^*, ..., m_N^*) \\ & + q(m_1^*, m_2, ..., m_N) - q(m_1^*, m_2^*, ..., m_N) + \frac{1}{N} q(m_1^*, ..., m_N^*) \\ & \cdots \\ & + q(m_1^*, ..., m_{N_1}^*, m_N) - q(m_1^*, ..., m_{N-1}^*, m_N^*) + \frac{1}{N} q(m_1^*, ..., m_N^*) \\ = \; & \sum_{i \in I} \pi_i^q(m_i). \end{aligned}$$

Therefore $(M_1, ..., M_N, q)$ is an additive mechanism. □

Proof of Proposition I.10

Proof. Let σ^* be any interim dominant strategy equilibrium of the mechanism. It follows that for player i, any type t_i where $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) > 0$ only sends messages that maximize $\pi_i^g(\cdot)$ with positive probability (and denote that maximized value as $\overline{\pi}_i$), and any type t_i where $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) < 0$ only sends messages that minimize $\pi_i^g(\cdot)$ with positive probability (and denote that minimized value as $\underline{\pi}_i$). For any t_i where $\alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) = 0$ let $\hat{\pi}_i(t_i)$ denote the expected value of $\pi_i(\cdot)$ conditional on t_i's (mixed) strategy under σ. Define $\lambda_i := \overline{\pi}_i - \underline{\pi}_i$ and

$$\mu_i(t_i) = \begin{cases} 1 & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) > 0 \\ 1 & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) = 0 \text{ and } \lambda_i = 0 \\ (\hat{\pi}_i(t_i) - \underline{\pi}_i)/\lambda_i & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) = 0 \text{ and } \lambda_i \neq 0 \\ 0 & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) < 0. \end{cases}$$

It is straightforward to verify that the induced social choice function q_σ satisfies $q_\sigma(t) = \sum_{i \in I} \lambda_i \mu_i(t_i) + \sum_{i \in I} \underline{\pi}_i$. Thus q_σ is a random dictatorship. □

Proof of Proposition I.13

Proof. Without loss, we assume $q(1, m_{-i}) \geq q(0, m_{-i})$. It is easy to check that the following strategy profile σ^* is an interim dominant strategy equilibrium,

$$\sigma_i^*(t_i) = \begin{cases} 1 & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) \geq 0 \\ 0 & \text{if } \alpha_i(\hat{\theta}_i(t_i), \hat{b}_i(t_i)) < 0. \end{cases}$$

□

Proof of Proposition I.14

Proof. The "only if" direction is obvious. For the "if" direction we prove its contrapositive: If a mechanism is not DSIC, then it cannot be both EPIC and IDSIC. Suppose, in order to lead to a contradiction, that there exists q that is not DSIC, yet it is EPIC and IDSIC. That q is not DSIC implies that

$$u_i(\theta)q(\theta_i, \theta'_{-i}) < u_i(\theta)q(\theta'_i, \theta'_{-i})$$

for some agent i, $\theta_i, \theta'_i \in \Theta_i$ and $\theta'_{-i} \in \Theta_{-i}$. Suppose $u_i(\theta) > 0$, then we have $q(\theta_i, \theta'_{-i}) < q(\theta'_i, \theta'_{-i})$, which implies that $\alpha_i(\theta_i) \leq 0$ and $\alpha_i(\theta'_i) \geq 0$ by Corollary I.8.

Suppose $\alpha_i(\theta_i) < 0$. It then follows from Proposition I.8 that $q(\theta_i, \theta_{-i}) \leq q(\theta'_i, \theta_{-i})$. If $q(\theta_i, \theta_{-i}) < q(\theta'_i, \theta_{-i})$ then $u_i(\theta)q(\theta_i, \theta_{-i}) < u_i(\theta)q(\theta'_i, \theta_{-i})$, contradicting EPIC. If $q(\theta_i, \theta_{-i}) = q(\theta'_i, \theta_{-i})$ then $q(\theta'_i, \theta_i) - q(\theta_i, \theta_{-i}) = 0$, which implies, by condition 1 of Proposition I.8, that $\min_{\hat{\theta}_{-i} \in \Theta_{-i}} \left(\overline{q}(\theta_i, \hat{\theta}_{-i}) - \underline{q}(\theta_i, \hat{\theta}_{-i}) \right) = q(\theta'_i, \theta_{-i}) - q(\theta_i, \theta_{-i}) = 0$. Also by condition 1 of Proposition I.8 we have $\max_{\hat{\theta}_{-i} \in \Theta_{-i}} \left(\overline{q}(\theta'_i, \hat{\theta}_{-i}) - \underline{q}(\theta_i, \hat{\theta}_{-i}) \right) \geq q(\theta'_i, \theta'_{-i}) - q(\theta_i, \theta'_{-i}) > 0$, therefore

$$\underline{\alpha}_i(\theta_i) \max_{\hat{\theta}_{-i} \in \Theta_{-i}} \left(\overline{q}_i(\hat{\theta}_{-i}) - \underline{q}_i(\hat{\theta}_{-i}) \right) + \overline{\alpha}_i(\theta_i) \min_{\hat{\theta}_{-i} \in \Theta_{-i}} \left(\overline{q}_i(\hat{\theta}_{-i}) - \underline{q}_i(\hat{\theta}_{-i}) \right) < 0$$

because $\alpha_i(\theta_i) < 0$ implies $\underline{\alpha}_i < 0$. This, however, contradicts condition 2 of Proposition I.8.

We can thus deduce that $\alpha_i(\theta_i) = 0$. That $u_i(\theta) > 0$ implies $\overline{\alpha}_i(\theta_i) > 0$, and hence $\underline{\alpha}_i(\theta_i) = -\overline{\alpha}_i(\theta_i) < 0$. Substituting this into inequality I.2, we have $\overline{\alpha}_i(\theta_i) \Big((q(\theta_i, \theta'_{-i}) - q(\theta'_i, \theta'_{-i})) - (q(\theta_i, \theta_{-i}) - q(\theta'_i, \theta_{-i})) \Big) = 0$, which implies that $q(\theta_i, \theta_{-i}) - q(\theta'_i, \theta_{-i}) = q(\theta_i, \theta'_{-i}) - q(\theta'_i, \theta'_{-i}) < 0$, contradicting EPIC because $u_i(\theta) > 0$.

If $u_i(\theta) < 0$ then similar contradictions arise analogously. □

CHAPTER II

Social Discounting and Intergenerational Pareto

The most critical issue in evaluating policies and projects that affect generations of individuals is the choice of social discount rate. This chapter shows that there exist social discount rates such that the planner can simultaneously be (i) an exponential discounting expected utility maximizer; (ii) intergenerationally Pareto—i.e., if all individuals from all generations prefer one policy/project to another, the planner agrees; and (iii) strongly non-dictatorial—i.e., no individual from any generation is ignored. Moreover, to satisfy (i)–(iii), if the time horizon is long enough, it is generically sufficient and necessary for social discounting to be more patient than the most patient individual's long-run discounting, independent of the social risk attitude.

II.1 Introduction

Many economic decisions are inherently dynamic and affect multiple generations, such as corporate and household long-term investment decisions, intertemporal taxation, durable public good provision, environmental policies, etc. These decisions crucially depend on one parameter, the *social discount rate*, which encapsulates the trade-off between the current benefit and future benefit from the society's point of view. Unfortunately, there is no consensus on which social discount rate should be used. This disagreement has sparked debate, for example, about the cost-benefit analysis of environmental projects that affect many, if not all, future generations. Moreover, the evaluation of those projects is sensitive to the choice of social discount rate. The famous Stern review uses a near-zero social discount rate (pure rate of time preference), and suggests that we should take strong and immediate action on climate change (see Stern (2007)).[1] Nordhaus (2007) argues that Stern's conclusion does

This chapter is based on the paper "Social Discounting and Intergenerational Pareto " (Feng and Ke, 2018).

[1]The consumption discount rate derived from the Ramsey formula used in the Stern review depends on the pure rate of time preference, the elasticity of the marginal utility of consumption, and the growth rate

not hold if a market rate is used instead. Many economists, however, believe that using a high discount rate (such as a market rate) is ethically indefensible.

In the social discounting literature, some economists have argued that social discounting should be more patient than individual discounting (for example, see Caplin and Leahy (2004) and Farhi and Werning (2007)). The idea is that if social discounting takes into account how future generations will feel about their consumption, then because future generations will value future consumption relatively more than the current generation values future consumption, social discounting will also value future consumption more than the current generation does.[2] However, these studies usually assume that only one (representative) individual is in the society. How their insight carries over to a society with heterogeneous individuals—and which individual's discounting social discounting should be more patient than—remains unanswered.

Let us explain what will go wrong with heterogeneous individuals. What is common among these dynamic economic decisions is that there is a benevolent planner who must make choices from risky alternatives for generations of individuals. In such a setting, first, economists often assume that the planner's objective is an *exponential discounting (expected) utility function*. This assumption is widely used and normatively appealing, because it is equivalent to assuming that the planner's preference is time-consistent, time-invariant, and stationary.[3] Second, it is often assumed that a benevolent planner respects individuals' preferences. In other words, some notion of the *Pareto* property should hold: If "all" individuals agree that one policy/project is better than another, the planner should agree that the former is better.

Despite the fact that these two assumptions are fundamental to economics, they cannot be satisfied simultaneously (see Gollier and Zeckhauser (2005), Zuber (2011), and Jackson and Yariv (2015)). Even if every individual has an exponential discounting utility function, a planner must be dictatorial to ensure that her exponential discounting utility function satisfies some Pareto property. The negative result also challenges the conclusion that social discounting should be more patient than individual discounting. In light of the negative result, with heterogeneous individuals, perhaps we can only conclude that the planner is more patient than the only individual (dictator) she cares about.

This paper addresses these issues using a classic approach. We introduce a new Pareto

of per capita consumption.

[2]Some economists have also argued that individuals' altruistic discounting for future generations should be excluded from the planner's aggregation. See Hammond (1987) and Boadway (2012).

[3]A version of the definition of time consistency, time invariance, and stationarity can be found in Halevy (2015). Under the assumption that the utility function is a time-additively separable expected utility function, Halevy's version of the three properties is equivalent to assuming an exponential discounting expected utility function.

property, and characterize the range of (pure-time-preference) social discount rates that are compatible with the new property. In models that generate the negative result, there is often only one generation of individuals. The Pareto property they use, which we call *current-generation Pareto*, is the key to the negative result. Current-generation Pareto requires that whenever a consumption sequence $\mathbf{p}$ is preferred to another sequence $\mathbf{q}$ by every current-generation individual, then the planner prefers $\mathbf{p}$ to $\mathbf{q}$. In many problems we consider, multiple generations of individuals are involved. As Pigou (1920) argues, the planner should not only respect how the current generation discounts the future, but also care about the actual well-being of future generations—that is, how future generations will feel about their consumption and how they will discount the future. The Pareto property we introduce, *intergenerational Pareto*, captures this. It requires that whenever a consumption sequence $\mathbf{p}$ is preferred to $\mathbf{q}$ by every individual from every generation, then the planner prefers $\mathbf{p}$ to $\mathbf{q}$.

Specifically, each generation-t individual i lives for one period, and has a discount function $\delta_i(\tau - t)$ to discount period-τ consumption.[4] The planner is intergenerationally Pareto and has an exponential discounting utility function. To contrast with the negative result, we require that the planner be *strongly non-dictatorial* in the sense that she never ignores the preference of any individual from any generation. Under these assumptions, we show how the range of social discount factors depends on (a) individual relative discounting, average discounting, and long-run discounting, and (b) the linear dependency of individual instantaneous utility functions.[5]

We first characterize the range of social discount factors assuming that individual discount functions are exponential. This allows us to compare our results to the negative result directly. We examine two cases. In the first case, individuals share the same instantaneous utility function. In this way, we focus on aggregating individual discount functions. The negative result is avoided: We find that the planner is intergenerationally Pareto and strongly non-dictatorial if and only if the social discount factor is higher than the *least patient* individual's discount factor.

Since the least patient individual's discount factor could be quite low, a wide range of social discount factors can be supported by the first result. The result will be rather different in our second case in which individual instantaneous utility functions are linearly independent. When there are many consumption goods, individual instantaneous utility functions are generically linearly independent. Under this assumption, we find that the planner is intergenerationally Pareto and strongly non-dictatorial if and only if the social

[4]Individuals *altruistically* care about future generations' consumption.

[5]The discount rate is equal to one minus the discount factor.

discount factor is higher than the *most patient* individual's discount factor, *independent of the planner's instantaneous utility function*. This result thus provides a new justification for the use of a near-zero social discount rate.

In general, individual discount functions are not exponential. One challenge that comes with general individual discount functions is that when we say that social discounting is more patient than individual discounting, it is not even clear what individual discounting refers to. We show that when individuals share the same instantaneous utility function, there exist two cutoffs for the social discount factor. One is related to the least patient individual's maximal relative discount factor, and the other to the least patient individual's asymptotic average discount factor. If the social discount factor is above the first cutoff, the planner is intergenerationally Pareto and strongly non-dictatorial. Conversely, if the social discount factor is below the second cutoff, the planner must violate intergenerational Pareto as long as the time horizon is long enough; that is, there exist two consumption sequences such that every individual from every generation thinks that one is better than the other, but the planner disagrees. The two cutoffs are tight.

The two cutoffs merge into one cutoff when individuals exhibit *present bias*. The unique cutoff is equal to the least patient individual's *long-run discount factor*. Each individual's long-run discount factor is defined as the asymptotic relative discount factor and the asymptotic average discount factor.

Lastly, if individual instantaneous utility functions are linearly independent, the cutoff for the social discount factor jumps from the least patient individual's long-run discount factor to the most patient one's, again independent of the planner's instantaneous utility function. We also characterize how the cutoff for the social discount factor changes gradually from the least patient individual's long-run discount factor to the most patient one's, as the number of types of individual instantaneous utility functions increases.

Related Literature

This paper is not the first to aggregate the preferences of multiple generations of individuals. Indeed, there is a long-running debate on whether future generations should be aggregated. Among others, Pigou (1920), Ramsey (1928), Sen (1961), Feldstein (1964), Solow (1974), Arrow (1999), Caplin and Leahy (2004), and Farhi and Werning (2007) are in favor, and Eckstein (1957), Bain (1960), and Marglin (1963) believe that the government's or the policy maker's decision should only reflect the preferences of the current generation. Our approach is closer to Caplin and Leahy and Farhi and Werning, who show that assuming there is only one individual in each generation, social discounting should be more patient than the sole

individual's discounting. Our results show that having multiple heterogeneous individuals in each generation makes an important difference.

Many papers have analyzed the aggregation of one generation of heterogeneous individuals. Weitzman (2001) conducts a survey of economists' discount rates to motivate a gamma discounting model. Gollier and Zeckhauser (2005) study a dynamic efficient allocation problem with heterogeneous individuals and show that even when individuals have constant discount rates, the representative agent has a decreasing discount rate. Zuber (2011) establishes that a planner cannot have an exponential discounting utility function and be (current-generation) Pareto when individuals have private consumption. Jackson and Yariv (2015) present a similar negative result, in which consumption is public. Millner and Heal (2018) show that the negative result goes away if we only require that the planner's objective be time-consistent. A key difference between these papers and ours is that they aggregate only one generation of individuals, whereas we aggregate multiple generations. This distinction is important in economic decisions that have long-term impact, such as environmental policies and intertemporal taxation.

There are other approaches to the study of social discounting. Our paper emphasizes the relation between social discounting and individual discounting implied by intergenerational Pareto. Chambers and Echenique (2018) study three models of discount rates. One aggregates exponential individual discount functions in a utilitarian way, which is similar to Weitzman (2001), Zuber (2011), and Jackson and Yariv (2014, 2015). The other two aggregate exponential individual discount functions by selecting the most pessimistic utilitarian weight and discount function, respectively. Millner (2020) shows that if heterogeneous individuals are not fully paternalistic, they will agree on parameters for the long-run social discount rate. Zuber and Asheim (2012), Asheim and Zuber (2014), Fleurbaey and Zuber (2015), and Piacquadio (2017) study models in which social discounting is due to intergenerational inequality aversion. Jonsson and Voorneveld (2018) study a welfare criterion for multiple generations. Each generation has one individual, and in the limit of the criterion, different generations are treated equally.

In the first part of Drugeon and Wigniolle (2017), they characterize what exponential discounting utility functions can be written as weighted sums of individuals' current selves' and future selves' quasi-hyperbolic discounting utility functions; their result is related to our Propositions II.4 and II.5, and Proposition 5 of Galperti and Strulovici (2017). Drugeon and Wigniolle (2016) and the second part of Drugeon and Wigniolle (2017) study time-consistent solutions for consumption-saving problems with heterogeneous exponential and quasi-hyperbolic discounting individuals, respectively. The planner in period t maximizes the weighted sum of period-t individuals' utility, and the solution is the subgame perfect

Nash equilibrium of the game between the planner's multiple selves.

Our paper is also related to Mongin (1998), who establishes that under a standard form of Pareto, as long as individuals' subjective probabilities are linearly independent or their instantaneous utility functions are affinely independent, the planner must be dictatorial. Related results can be found in Mongin (1995) and Chambers and Hayashi (2006). In our model, if we view periods as states and discount factors as subjective probabilities, Mongin's result seems to apply. Nonetheless, our planner is not dictatorial. The technical reason why our Theorem 6 can bypass Mongin's negative result is the assumption that all individuals share the same instantaneous utility function. As for Theorem 9, we first aggregate individual utility functions with identical instantaneous utility functions into an EDU function whose discount factor is equal to the social discount factor. Then, we aggregate utility functions with identical discount factors (subjective probabilities). Both steps bypass Mongin's negative result.

The paper proceeds as follows. In Section 2, we describe individuals' and the planner's preferences. Section 3 introduces a variant of the negative result and two key assumptions of the paper, intergenerational Pareto and strong non-dictatorship. We characterize the range of social discount factors under the assumption that individuals have exponential discount functions in Section 4, and under the assumption that individuals have general discount functions in Section 5. Section 6 concludes.

II.2 Preferences

There are $2 < T < +\infty$ generations/periods. In each generation, $1 < N < +\infty$ individuals live for one period. With an abuse of notation, let $N := \{1, \ldots, N\}$ and $T := \{1, \ldots, T\}$. The generation-t individual i is the parent of the generation-$(t+1)$ individual i, in which $t, t+1 \in T$ and $i \in N$. In each period, there is a public risky consumption good denoted by $\Delta(X)$, in which $\Delta(X)$ is the set of probability measures on a compact set $X \subset \mathbb{R}^m$.[6] A typical consumption sequence is denoted by $\mathbf{p} = (p_1, \ldots, p_T) \in \Delta(X)^T$.[7]

Although individuals live for one period, they altruistically care about future generations' consumption. We assume throughout the paper that the generation-t individual i has

[6] All results we derive apply to the case in which each individual has his own consumption. We only need to view public consumption as an N-tuple of individual consumption, and let each individual care only about his own component.

[7] We discuss what may change if we allow uncertainty to resolve over time in Section II.8.4 in the Supplemental Material.

the following *discounting utility* function:

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_i(\tau - t) u_i(p_\tau), \tag{II.1}$$

in which $\delta_i : \{0, \ldots, T-1\} \to \mathbb{R}_{++}$ with $\delta_i(0) = 1$ is called the *discount function*, and the *instantaneous utility function* $u_i : \Delta(X) \to \mathbb{R}$ is a continuous expected utility function. The generation-t individual i's discounting utility function induces a preference, denoted by $\succsim_{i,t}$, over consumption sequences $\Delta(X)^T$.

We have assumed that the generation-$(t+1)$ individual i inherits the generation-t individual i's discount function and instantaneous utility function. This assumption does not imply that a parent and his offspring have the same preference, because the generation-$(t+1)$ individuals' discount functions are shifted one period forward. This assumption simplifies our analysis and can be relaxed (see Section II.8.1 in the Supplemental Material).

In each period $t \in T$, the planner's objective is an *exponential discounting utility* (EDU) function:

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u(p_\tau), \tag{II.2}$$

in which $\delta > 0$ is the *social discount factor*, and u, a continuous expected utility function on $\Delta(X)$, is the planner's instantaneous utility function. In each period $t \in T$, U_t induces the planner's preference, denoted by $\succsim_t$, over consumption sequences $\Delta(X)^T$.

It is well known that if the planner's objective is a discounting utility function, the planner is time-consistent if and only if the planner's discount function is exponential.[8] More generally, (II.2) holds if and only if the planner's preference is time-consistent, time-invariant, and stationary (see footnote 3). Also note that (II.2) holds for every $t \in T$; that is, the social discount factor and the planner's instantaneous utility function never change.

Lastly, to rule out uninteresting cases and simplify the statement of our results, we assume that there are some fixed consequences $x_*, x^* \in X$ such that for any $i \in N$, $u_i(x_*) = u(x_*) = 0$ and $u_i(x^*) = u(x^*) = 1$ throughout the paper. A similar assumption, called the *minimum agreement condition*, also appears in De Meyer and Mongin (1995). Our main findings do not rely on this assumption, and we provide a more detailed discussion following Lemma 3. More generally, for any continuous expected utility function v defined on $\Delta(X)$, we say that it is *normalized* if $v(x^*) = 1$ and $v(x_*) = 0$. One may think of x^* as one dollar and x_* as zero dollars, or x^* as the best consumption good and x_* as the worst.

[8] Since individuals only live for one period, time consistency may have a nonstandard interpretation for them. In contrast, the planner is a long-lived entity who tries to stick to an objective function that exhibits nice properties. The interpretation of time consistency for the planner is similar to the standard one.

II.3 Intergenerational Pareto and Dictatorship

We want to assume that the planner's preference $(\succsim_t)_{t\in T}$ satisfies some Pareto property. In a dynamic setting, however, there are multiple ways to define the Pareto property. Different notions of Pareto have different implications. For example, Zuber (2011) and Jackson and Yariv (2015) show that if a planner has an EDU function and follows their Pareto property, the planner must be dictatorial. To motivate our new Pareto property, it is useful to first understand the negative result. Below, we introduce a version of the negative result.

II.3.1 A Variant of the Negative Result

Below is a variant of the Pareto property used by Zuber (2011) and Jackson and Yariv (2015) that fits our setting.

Definition 1. The planner's preference $(\succsim_t)_{t\in T}$ is current-generation Pareto if for any consumption sequences $\mathbf{p}, \mathbf{q} \in \Delta(X)^T$, in each period $t \in T$, $\mathbf{p} \succsim_{i,t} \mathbf{q}$ for all $i \in N$ implies $\mathbf{p} \succsim_t \mathbf{q}$, and $\mathbf{p} \succ_{i,t} \mathbf{q}$ for all $i \in N$ implies $\mathbf{p} \succ_t \mathbf{q}$.

This notion of Pareto says that in any period t, if all current-generation individuals agree that a consumption sequence $\mathbf{p}$ is preferred to another sequence $\mathbf{q}$, the planner should agree that $\mathbf{p} \succsim_t \mathbf{q}$. The same applies when the preferences are all strict.

Consider a situation in which every generation-t individual i has an EDU function; that is, for some discount factor $\delta_i > 0$,

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_i^{\tau-t} u_i(p_\tau).$$

Let us present below a variant of the negative result.

Proposition II.1. *Suppose each generation-t individual i has an EDU function with discount factor δ_i and instantaneous utility function u_i such that δ_i's are distinct. The planner is current-generation Pareto if and only if there exists some $i \in N$ such that for any $t \in T$, $U_t = U_{i,t}$.*

The result says that if we require that the planner be current-generation Pareto and have an EDU function, the planner's preference must be identical to some individual's preference in every period. Since consumption is public, our setting is closer to Jackson and Yariv (2015). However, Jackson and Yariv's result is different from the above proposition. For example, they require that instantaneous utility functions be defined on a one-dimensional

space and twice continuously differentiable, and we require that instantaneous utility functions be expected utility functions.

The intuition is as follows. First, the planner is current-generation Pareto if and only if her EDU function is equal to a weighted sum of the individuals' EDU functions; because we consider expected utility functions, this is an implication of Harsanyi (1955). Next, for simplicity, suppose there are only two individuals with identical instantaneous utility functions $u_1 = u_2$. The planner attaches a weight $\omega \in [0, 1]$ to the first individual and $1 - \omega$ to the second individual. Now, for the planner to not be dictatorial, there must be some $\omega \in (0, 1)$ and $\delta > 0$ such that

$$\omega\delta_1 + (1 - \omega)\delta_2 = \delta,$$

and

$$\omega\delta_1^2 + (1 - \omega)\delta_2^2 = \delta^2.$$

However, one cannot find such a δ, unless $\omega = 0$ or 1.

II.3.2 Intergenerational Pareto

A key feature of environmental policies and many other economic policies is that such decisions affect multiple generations. Current-generation Pareto only takes into account the preferences of the current generation. Although current-generation individuals altruistically care about future consumption and the planner should respect how they discount the future, how they think about the future may well differ from how future generations will think. Since future generations will be affected by the planner's decision, the planner should take into account their actual well-being (including how they will discount their own future). The following Pareto property captures this idea.

Definition 2. The planner's preference $(\succsim_t)_{t\in T}$ is intergenerationally Pareto if for any consumption sequences $\mathbf{p}, \mathbf{q} \in \Delta(X)^T$, in each period $t \in T$, $\mathbf{p} \succsim_{i,s} \mathbf{q}$ for all $i \in N$ and all $s \geq t$ implies $\mathbf{p} \succsim_t \mathbf{q}$, and $\mathbf{p} \succ_{i,s} \mathbf{q}$ for all $i \in N$ and all $s \geq t$ implies $\mathbf{p} \succ_t \mathbf{q}$.

Intergenerational Pareto says that in any period t, if all current- and future-generation individuals agree that a consumption sequence is preferred to another sequence, the planner should agree. For example, suppose all current-generation individuals are extremely selfish: They are willing to sacrifice the environment to increase their own consumption. If the planner is current-generation Pareto, the planner must agree with them, and let them destroy the environment. However, if the planner is intergenerationally Pareto, the planner is allowed to disagree with them, because what they prefer hurts future generations. Note that if

the planner is current-generation Pareto, she is also intergenerationally Pareto. Therefore, intergenerational Pareto is weaker than current-generation Pareto.

Our model considers expected utility functions. This enables us to apply the classic result from Harsanyi (1955) and Fishburn (1984) to characterize the consequence of intergenerational Pareto.

Lemma 3. *(Harsanyi (1955)) The planner's preference $(\succsim_t)_{t\in T}$ is intergenerationally Pareto if and only if in each period $t \in T$, there exists a finite sequence of nonnegative numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that*

$$U_t = \sum_{i=1}^{N}\sum_{s=t}^{T} \omega_t(i,s) U_{i,s}.$$

The lemma above follows from Harsanyi (1955) and Fishburn (1984), and shows that intergenerational Pareto is equivalent to intergenerational utilitarianism in our setting; that is, the planner is intergenerationally Pareto if and only if in each period, her utility function is equal to a weighted sum of all the current- and future-generation individuals' utility functions. In contrast, current-generation Pareto is equivalent to current-generation utilitarianism. We omit the proof of this lemma.

Instantaneous utility functions are normalized. In general, it is possible that there do not exist two consumption sequences such that all individuals strictly prefer one to the other; in that case, if the planner is indifferent to all consumption sequences, the planner will be intergenerational Pareto trivially. If the planner is always indifferent, her instantaneous utility function is constant and her discount function can be arbitrary. The normalization assumption rules out this uninteresting case.

II.3.3 Dictatorship

In the negative result, a planner can have an EDU function and be current-generation Pareto as long as she is dictatorial. To rule out dictatorship, we introduce a strong notion of non-dictatorship such that not only is the planner not dictatorial, but also every individual from every generation has a say.[9]

Definition 4. We say that the planner is strongly non-dictatorial if for each $t \in T$,

$$U_t(\mathbf{p}) = f_t\left(U_{1,t}(\mathbf{p}), \ldots, U_{1,T}(\mathbf{p}), U_{2,t}(\mathbf{p}), \ldots, U_{2,T}(\mathbf{p}), \ldots, U_{N,T}(\mathbf{p})\right)$$

for some (strictly) increasing function f_t.

[9]When the planner is not dictatorial, we only know that at least two individuals' preferences are taken into account by the planner.

In light of Lemma 3, under intergenerational Pareto, this means that the planner's utility function can be written as a weighted sum of individual utility functions with positive weights.

Intergenerational Pareto is weaker than current-generation Pareto. According to Lemma 3, the planner has more utilitarian weights to assign under intergenerational Pareto, which makes it easier for the planner to aggregate individuals' utility functions into an EDU function. The strongly non-dictatorial property, on the other hand, makes the aggregation problem harder, because it requires that all weights be positive.

II.4 Individuals with Exponential Discount Functions

We address two aspects of social discounting. First, can we bypass the negative result? If so, which social discount factors are reasonable? In particular, which social discount factors, under our assumptions, are compatible with intergenerational Pareto? Second, recall that in the social discounting literature, economists have argued that the social discount factor should be higher than the individual discount factor. Accordingly, with heterogeneous individuals, which individual's discount factor should the social discount factor be higher than?

To contrast with the negative result, we first examine a special case of our model in which individual discount functions are exponential.

II.4.1 Aggregating Individual Discount Functions

To focus on discounting, suppose that all individual instantaneous utility functions are identical; that is, there is some continuous expected utility function $u : \Delta(X) \to \mathbb{R}$ such that each generation-t individual i's utility function is

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_i^{\tau-t} u(p_\tau).$$

This assumption will be relaxed soon, and we will use the result established under this assumption to highlight how the range of reasonable social discount factors is affected by individual instantaneous utility functions. An alternative interpretation of this assumption is that the planner only wants to aggregate individual discount functions. Therefore, it is without loss of generality to replace the (possibly heterogeneous) individual instantaneous

utility functions with the planner's instantaneous utility function u.[10]

Proposition II.2. *Suppose each generation-t individual i has an EDU function with discount factor δ_i and instantaneous utility function u. Let the planner's instantaneous utility function be u. The planner is intergenerationally Pareto and strongly non-dictatorial if and only if $\delta > \min_i \delta_i$.*

When individuals share the same instantaneous utility function, according to Lemma 3, the planner must use the same instantaneous utility function in order to satisfy the Pareto property.

Proposition II.2 shows that under intergenerational Pareto rather than current-generation Pareto, a positive result can be established. Moreover, under the current set of assumptions, it is the least patient individual's discount factor that the social discount factor should be higher than.

Because discount functions are exponential and consumption is public, Proposition II.2 can be directly compared to Jackson and Yariv (2014, 2015). In Jackson and Yariv, adding more current-generation exponential discounting individuals to the aggregation cannot help eliminate the negative result. In contrast, we add future-generation exponential discounting individuals to the aggregation, and this helps.

To see why, first recall that when $u_i = u$, Jackson and Yariv (2014) show that utilitarian aggregation of the current generation leads to a social discount function that exhibits present bias. The fact that future generations will not care about past consumption as much as past generations did helps us remove the present bias. In our model, past consumption does not enter future generations' utility functions; that is, $\delta_i(\tau) = 0$ for any $\tau < 0$. This implies that, for example, generation-t individual i's relative discount factor applied to period-t consumption (relative to period-$(t-1)$ consumption) is equal to "$\delta_i(0)/\delta_i(-1) = +\infty$." Thus, generation-$t$ is "infinitely patient" between period $t-1$ and period t. The infinite patience can be used in the aggregation to offset the present bias generated by aggregating the current generation alone. In fact, the same result continues to hold even if individuals backward discount past consumption exponentially (see Section II.8.3 in the Supplemental Material).[11]

To understand how the proposition is proved, consider the planner in the first period. According to Lemma 3, the planner is intergenerationally Pareto if and only if her objective

[10] In this interpretation, however, each individual i's preference in the definition of Pareto properties must be replaced with another preference induced by a discounting utility function with a discount function δ_i and an instantaneous utility function u chosen by the planner.

[11] See Caplin and Leahy (2004) and Ray, Vellodi and Wang (2017) for models that allow backward discounting for past consumption.

function U_1 satisfies

$$U_1(\mathbf{p}) = \sum_{t=1}^{T} \sum_{i=1}^{N} \omega(i,t) U_{i,t}(\mathbf{p}), \tag{II.3}$$

for any consumption sequence $\mathbf{p}$, in which $\omega(i,t) \geq 0$ is the weight the planner assigns to the generation-t individual i. Consider how the planner discounts period-τ consumption. Since the planner and individuals have EDU functions, and their instantaneous utility functions are identical, equation (II.3) becomes

$$\delta^{\tau-1} = \sum_{t=1}^{\tau} \sum_{i=1}^{N} \omega(i,t) \delta_i^{\tau-t}. \tag{II.4}$$

To prove the "if" part of the proposition, we let all individuals' weights be equal to some small numbers, except for the least patient individuals. We show that if those weights are sufficiently small, there exist positive weights for the least patient individuals such that the weighted sum of all individuals' utility functions is an EDU function with the social discount factor $\delta > \min_i \delta_i$. For example, suppose $N = T = 2$ and $\delta_1 < \delta_2$.[12] Let $\omega(i,t) = \varepsilon$ whenever $i = 2$. Equation (II.4) implies that

$$\omega(1,1) = 1 - \omega(2,1) = 1 - \varepsilon$$

and

$$\omega(1,2) = \delta - \omega(1,1)\delta_1 - \omega(2,1)\delta_2 - \omega(2,2) = \delta - \delta_1 + \varepsilon(\delta_1 - \delta_2 - 1).$$

Since $\delta > \delta_1$, $\omega(1,1)$ and $\omega(1,2)$ are positive when $\varepsilon = 0$. Therefore, when $\varepsilon > 0$ is sufficiently small, $\omega(i,t)$'s can all be positive.

To understand the "only-if" part, suppose individual 1's discount factor is the lowest. By letting $\tau = 1$, equation (II.4) implies that $\sum_{i=1}^{N} \omega(i,1) = 1$. Since the planner is strongly non-dictatorial, we can assume that $\omega(i,t)$'s are positive, and hence $\sum_{t=1}^{\tau} \sum_{i=1}^{N} \omega(i,t) > 1$. Then, equation (II.4) implies that

$$\delta^{\tau-1} = \sum_{t=1}^{\tau} \sum_{i=1}^{N} \omega(i,t) \delta_i^{\tau-t} \geq \delta_1^{\tau-1} \sum_{t=1}^{\tau} \sum_{i=1}^{N} \omega(i,t) > \delta_1^{\tau-1}, \tag{II.5}$$

which means $\delta > \delta_1$.

Note that Proposition II.2 is not very helpful in pinning down social discount factors, because the least patient individual's discount factor can be quite low. Thus, many social

[12] We have assumed $T > 2$ because when $T \leq 2$, there will be no negative result (such as Proposition II.1) trivially. However, to illustrate the idea of the proof here, we only need an example with $T = 2$.

discount factors can satisfy our requirements. However, as will be shown below, this is no longer the case once we relax the unrealistic assumption that individuals share the same instantaneous utility function.

II.4.2 Social Discounting and Individual Instantaneous Utility Functions

The assumption that individuals share the same instantaneous utility function is clearly unrealistic. As long as $|X| \geq N$ (i.e., the number of deterministic consumption goods is higher than the number of individuals in each generation), generically, the instantaneous utility functions should not only be different, but also linearly independent.[13]

Definition 5. An N-tuple of continuous expected utility functions $(u_i)_{i\in N}$ is linearly independent if there are no constants $\alpha_1, \ldots, \alpha_N$ that are not all zero, and $\sum_{i\in N} \alpha_i u_i(p) = 0$ for all $p \in \Delta(X)$.

It turns out that when individual instantaneous utility functions are linearly independent, the cutoff for the social discount factor jumps from $\min_i \delta_i$ to $\max_i \delta_i$; that is, generically, the social discount factor should be higher than the *most* patient individual's discount factor.

Proposition II.3. *Suppose each generation-t individual i has an EDU function with discount factor δ_i and instantaneous utility function u_i such that $(u_i)_{i\in N}$ is linearly independent. Let the planner's instantaneous utility function u be an arbitrary strict convex combination of $(u_i)_{i\in N}$.*[14] *The planner is intergenerationally Pareto and strongly non-dictatorial if and only if $\delta > \max_i \delta_i$.*

To understand why we assume that the planner's instantaneous utility function is a strict convex combination of individual instantaneous utility functions, note that Lemma 3 implies that the intergenerationally Pareto and strongly non-dictatorial planner's utility function is equal to a weighted sum of individual discounting utility functions with positive weights. Thus, the planner's instantaneous utility function must also be a positively weighted sum of individual instantaneous utility functions. Since instantaneous utility functions are normalized, the weights sum up to 1.

[13] However, for example, if individual instantaneous utility functions are drawn from some fixed small set of continuous expected utility functions rather than the set of all continuous expected utility functions, or the number of consumption goods is lower than N, individual instantaneous utility functions need not be linearly independent. See Theorem 10 for results without assuming linear independence.

[14] By a strict convex combination of $(u_i)_{i\in N}$, we mean that u is in the interior of the convex hull of $u_1, \ldots, u_N$.

Notice that the planner's instantaneous utility function—in other words, her risk attitude—is independent of the cutoff for the social discount factor. This is somewhat surprising. Suppose there are two individuals, 1 and 2, and individual 2 is more patient. The above result says that even if the social discount factor is close to individual 2's discount factor, it is not necessarily the case that the planner's risk attitude is also close to individual 2's risk attitude. We can have a planner whose risk attitude is close to individual 1's, but the social discount factor is close to individual 2's.

If there are many individuals with a wide range of discount factors, this result may imply that the planner must be very patient in order to be intergenerationally Pareto and strongly non-dictatorial. This provides a new justification for the use of the near-zero social discount rate by Stern (2007). If one thinks that a market rate is higher than the lowest individual discount rate, this result also rules out the use of a market rate as the social discount rate.

This result shows that the cutoff for the social discount factor in Proposition II.2 is not robust. When $u_i = u_j$ for any $i, j \in N$, the cutoff is $\min_i \delta_i$. If we introduce a small perturbation to u_i's, generically, the cutoff jumps discontinuously to $\max_i \delta_i$.

One may wonder whether there is any intermediate case that yields a cutoff for the social discount factor between $\min_i \delta_i$ and $\max_i \delta_i$. In Section II.5.4, under a more general assumption about individual instantaneous utility functions, we explain the intermediate cases.

To understand how this proposition is proved, consider again the planner in the first period. To prove the "if" part of Proposition II.3, we want to find positive weights $\omega(i,t)$'s such that the weighted sum of all individuals' EDU functions is equal to the planner's EDU function. Focus on one arbitrary $j \in N$. We show that we can find positive weights $\tilde{\omega}(j,1), \ldots, \tilde{\omega}(j,T)$ such that $\sum_{t \in T} \tilde{\omega}(j,t) U_{j,t}$ is equal to an EDU function with any discount factor that is higher than δ_j. In particular, we can find positive weights $\tilde{\omega}(i,1), \ldots, \tilde{\omega}(i,T)$ for each $i \in N$ such that

$$\sum_{t=1}^{T} \tilde{\omega}(i,t) U_{i,t}(\mathbf{p}) = \sum_{\tau=1}^{T} \delta^{\tau-1} u_i(p_\tau),$$

for any consumption sequence $\mathbf{p}$, because $\delta > \max_i \delta_i$. Now, since the planner's instantaneous utility function $u = \sum_{i \in N} \lambda_i u_i$ for some positive numbers λ_i's, we only need to let $\omega(i,t) = \lambda_i \tilde{\omega}(i,t)$.

The "only-if" part of Proposition II.3 may be surprising. Note that when $(u_i)_{i \in N}$ is linearly independent and u is in the interior of $\mathrm{co}(\{u_i\}_{i \in N})$, there is a unique way to write u as a strict convex combination of $(u_i)_{i \in N}$.[15] Suppose $\sum_{i \in N} \lambda_i u_i = u$ and $\sum_{i \in N} \lambda_i = 1$ for

[15] We use $\mathrm{co}(\cdot)$ to denote the convex hull of a set.

some positive numbers λ_i's. The planner's period-1 EDU function satisfies

$$U_1(\mathbf{p}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\omega(i,t)U_{i,t}(\mathbf{p}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\omega(i,t)\sum_{\tau=t}^{T}\delta_i^{\tau-t}u_i(p_\tau), \tag{II.6}$$

in which $\omega(i,t) > 0$ is the weight the planner assigns to the generation-t individual i. This implies that the planner's instantaneous utility function for period-1 consumption satisfies

$$u(p_1) = \sum_{i=1}^{N}\omega(i,1)u_i(p_1)$$

for any p_1. Because $u = \sum_{i\in N}\lambda_i u_i$ and $(u_i)_{i\in N}$ is linearly independent,

$$\omega(i,1) = \lambda_i \tag{II.7}$$

must hold for any $i \in N$. Similarly, for period-2 consumption, equation (II.6) implies that

$$\delta u(p_2) = \sum_{i=1}^{N}[\omega(i,1)\delta_i + \omega(i,2)]u_i(p_2)$$

for any p_2. Since instantaneous utility functions do not change over time, the unique way to write u as a strict convex combination of $(u_i)_{i\in N}$ does not change; that is, $\delta u(p_2) = \delta\sum_{i\in N}\lambda_i u_i(p_2)$ for any p_2. Then, for any $i \in N$,

$$\lambda_i\delta = \omega(i,1)\delta_i + \omega(i,2). \tag{II.8}$$

Equations (II.7) and (II.8), together with the strongly non-dictatorial property, imply that $\delta > \delta_i$ for any $i \in N$. Hence, $\delta > \max_i \delta_i$.

II.5 Individuals with General Discount Functions

Individual discount functions are often not exponential (see Strotz (1955), Laibson (1997), and Frederick, Loewenstein and O'Donoghue (2002)).[16] Allowing individuals to have general discount functions, as in (II.1), raises a challenge to our previous findings: When we say that the social discount factor should be higher than some individual's discount factor, it is not clear how individual discount factors should be defined. Our analysis below shows how

[16]In contrast, for normative reasons, we may prefer to require that the planner have an EDU function. Such a requirement also makes our positive results sharper.

the range of reasonable social discount factors depends on some asymptotic characteristic of general individual discount functions, and how our positive results in Section 4 can be generalized.

II.5.1 Aggregating Individual Discount Functions

Again, we begin with the case in which individual instantaneous utility functions are identical; that is, there is some continuous expected utility function $u : \Delta(X) \to \mathbb{R}$ such that each generation-t individual i's utility function is

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_i(\tau - t)u(p_\tau).$$

Because we will need to vary T in part of the results below, we assume that individual discount functions are well defined for natural numbers. Starting from a set of individual discount functions δ_i's defined over natural numbers $\mathbb{N}$, whenever a finite T is chosen, we restrict the domain of δ_i's to $\{0, \ldots, T-1\}$. For instance, suppose individuals have quasi-hyperbolic discount functions. We first define $\delta_i(\tau) = \beta_i \delta_i^{\tau-1}$ for any $\tau > 0$. Then, we choose T and focus on $\delta_i(0), \ldots, \delta_i(T-1)$.

For each individual discount function δ_i, we call $\sqrt[\tau]{\delta_i(\tau)}$ the *average discount function*, and $\frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ the *relative discount function*. The average discount function measures the equivalent exponential discount factor for τ-period-ahead consumption. The relative discount function captures the instantaneous discounting for consumption that is $\tau + 1$ periods ahead relative to consumption that is τ periods ahead.

We make two assumptions about the individual discount functions. The first assumption says that average discounting has a limit; that is,

$$\lim_{\tau \to \infty} \sqrt[\tau]{\delta_i(\tau)} \text{ exists.} \tag{A1}$$

The second assumption says that the relative discount function is bounded; that is,

$$\text{there exists some } \alpha > 0 \text{ such that } \frac{\delta_i(\tau+1)}{\delta_i(\tau)} < \alpha \text{ for all } \tau \geq 0. \tag{A2}$$

The following theorem characterizes the set of social discount factors that are compatible with intergenerational Pareto under these assumptions.

Theorem 6. *Suppose each generation-t individual i's discounting utility function has an instantaneous utility function u and a discount function δ_i such that (A1) and (A2) hold.*

Let the planner's instantaneous utility function be u. Then,

1. *for each $\delta > \min_i \max_{\tau\in\{0,\dots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$, the planner is intergenerationally Pareto and strongly non-dictatorial;*

2. *for each $\delta < \min_i \lim_{\tau\to\infty} \sqrt[\tau]{\delta_i(\tau)}$, there exists some $T^* > 0$ such that if $T \geq T^*$, the planner is not intergenerationally Pareto.*

The theorem shows how social discounting depends on individual discounting when individuals have heterogeneous general discount functions. We can find two cutoffs for the social discount factor. If the social discount factor is above the *least patient* individual's maximal relative discount factor, the planner's preference must be intergenerationally Pareto and strongly non-dictatorial. If the social discount factor is below the *least patient* individual's asymptotic average discount factor, the planner's preference must have violated the intergenerationally Pareto property *as long as T is large enough*.

The first part of the theorem confirms that positive results can still be established when individuals have arbitrary discount functions. Given a social discount factor, we can also apply this result to check whether intergenerational Pareto holds. The second part of the theorem says that if the social discount factor is too low, there must be two consumption sequences such that all individuals from all generations prefer one over the other, but the planner disagrees. A reasonable social discount factor should not allow this to happen.

Note that for any fixed T, $\max_{\tau\in\{0,\dots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)} \geq \sqrt[T-1]{\delta_i(T-1)}$, because

$$\sqrt[T-1]{\delta_i(T-1)} = \sqrt[T-1]{\frac{\delta_i(T-1)}{\delta_i(T-2)} \cdot \dots \cdot \frac{\delta_i(1)}{\delta_i(0)}}; \tag{II.9}$$

that is, $\sqrt[T-1]{\delta_i(T-1)}$ is the geometric mean of $\frac{\delta_i(\tau+1)}{\delta_i(\tau)}$'s. Therefore, $\max_{\tau\in\{0,\dots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ will not be lower than $\lim_{\tau\to\infty} \sqrt[\tau]{\delta_i(\tau)}$ when T is large enough, and hence the first cutoff for the social discount factor will eventually be higher than the second cutoff.

Although the first cutoff may be strictly higher than the second, the two cutoffs in the theorem are "tight" in the following sense. For the first cutoff, there exist some individual discount functions δ_i's and T such that if the social discount factor $\delta \leq \min_i \max_{\tau\in\{0,\dots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$, the planner cannot be both intergenerationally Pareto and strongly non-dictatorial. This happens, for example, when δ_i's are exponential. Similarly, for the second cutoff, we can find individual discount functions δ_i's such that for any finite T, if the social discount factor $\delta \geq \min_i \lim_{\tau\to\infty} \sqrt[\tau]{\delta_i(\tau)}$, the planner is intergenerationally Pareto. This happens, for example, when δ_i's are quasi-hyperbolic (see Section II.8.2 in the Supplemental Material).

To prove the first part of this theorem, we first focus on one arbitrary $i \in N$. The key step is to show that for each $t \in T$, we can find positive weights $(\tilde{\omega}(i,t,s))_{s=t}^{T}$ such that $\sum_{s=t}^{T} \tilde{\omega}(i,t,s)U_{i,s}$ is equal to an EDU function with any discount factor that is higher than $\max_{\tau\in\{0,\ldots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$. Therefore, the planner can use utilitarian aggregation to transform each generation-t individual i's discounting utility function into an EDU function with a discount factor that is higher than $\max_{\tau\in\{0,\ldots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$. Then, we only need to let the planner use utilitarian weights to aggregate these EDU functions, as in Proposition II.2.

The proof of the second part is similar to that of Proposition II.2. Suppose individual 1's asymptotic average discount factor is the lowest strictly. When τ is large enough (and hence T must be large enough), we know that $\delta_i(\tau-s) \geq \delta_1(\tau-s)$. Hence, (II.5) becomes

$$\delta^{\tau-1} = \sum_{t=1}^{\tau}\sum_{i=1}^{N} \omega(i,t)\delta_i(\tau-t) \geq \delta_1(\tau-1)\sum_{t=1}^{\tau}\sum_{i=1}^{N} \omega(i,t) \geq \delta_1(\tau-1).$$

Therefore, $\delta \geq \lim_{\tau\to\infty} \sqrt[\tau]{\delta_1(\tau)}$ when τ is large enough.

II.5.2 Individual Long-Run Discounting

It turns out that for many widely used classes of individual discount functions, the two cutoffs in Theorem 6 merge into one. This is not a coincidence, and will help us identify an important characteristic of the individual discount function that determines the cutoff for the social discount factor. Let us introduce the following assumption:

$$\text{the relative discount function } \frac{\delta_i(\tau+1)}{\delta_i(\tau)} \text{ is nondecreasing in } \tau. \tag{A3}$$

In the literature of time inconsistency, when an individual has a nondecreasing relative discount function, the individual has (weak) *present bias*. A discount function δ_i is hyperbolic if for some $\alpha_i, \beta_i > 0$, $\delta_i(\tau) = (1+\alpha_i\tau)^{-\beta_i}$, and is quasi-hyperbolic if for some $\beta_i \in (0,1]$ and $\delta_i > 0$, $\delta_i(\tau) = \beta_i\delta_i^{\tau}$ for any $\tau > 0$. Exponential, hyperbolic, and quasi-hyperbolic discount functions all satisfy (A3).

Under (A2) and (A3), $\frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ is nondecreasing and bounded. Hence, $\lim_{\tau\to\infty} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ exists. Whenever $\lim_{\tau\to\infty} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ exists, because the average discount factor is the geometric mean of relative discount factors (see equation (II.9)), the average discount factor also has a limit. Therefore, assumptions (A2) and (A3) imply (A1).

More importantly, when $\lim_{\tau\to\infty} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ exists, the asymptotic relative discount factor

and the asymptotic average discount factor coincide:

$$\lim_{\tau\to\infty}\frac{\delta_i(\tau+1)}{\delta_i(\tau)}=\lim_{\tau\to\infty}\sqrt[\tau]{\delta_i(\tau)}.$$

Definition 7. When $\lim_{\tau\to\infty}\frac{\delta_i(\tau+1)}{\delta_i(\tau)}$ exists, we call $\delta_i^*:=\lim_{\tau\to\infty}\frac{\delta_i(\tau+1)}{\delta_i(\tau)}=\lim_{\tau\to\infty}\sqrt[\tau]{\delta_i(\tau)}$ individual i's *long-run discount factor*.

The following corollary of Theorem 6 has only one cutoff for the social discount factor, and shows how social discounting is related to individual long-run discounting.

Corollary 8. *Suppose each generation-t individual i's discounting utility function has an instantaneous utility function u and a discount function δ_i such that (A2) and (A3) hold. Let the planner's instantaneous utility function be u. Then,*

1. *for each $\delta>\min_i\delta_i^*$, the planner is intergenerationally Pareto and strongly non-dictatorial;*
2. *for each $\delta<\min_i\delta_i^*$, there exists some $T^*>0$ such that if $T\geq T^*$, the planner is not intergenerationally Pareto.*

If individuals have hyperbolic discount functions, the cutoff for the social discount factor is $\min_i\delta_i^*=1$; if individuals have quasi-hyperbolic discount functions with $\delta_i(\tau)=\beta_i\delta_i^\tau$, the cutoff is $\min_i\delta_i^*=\min_i\delta_i$.

In Section II.8.2 in the Supplemental Material, we reinterpret the generation-$(t+s)$ individual i (with $s>0$) as a future self of the generation-t individual i, which offers a new interpretation of intergenerational Pareto and allows us to discuss how our findings are related to the time-inconsistency literature. In addition, we provide a stronger result similar to Corollary 8 for the case in which individuals have quasi-hyperbolic discount functions.

From here on, to simplify the statement of our results, we focus on the case in which long-run discount factors δ_i^*'s are well defined.

II.5.3 Social Discounting and Individual Instantaneous Utility Functions

Corollary 8 shows that if all individuals share the same instantaneous utility function, the social discount factor only has to be higher than the lowest individual long-run discount factor. As one may expect, when individual instantaneous utility functions are linearly independent, the cutoff for the social discount factor jumps from $\min_i\delta_i^*$ to $\max_i\delta_i^*$. Thus,

generically, if social discounting is more patient than the *most* patient individual's long-run discounting, the planner is intergenerationally Pareto and strongly non-dictatorial; otherwise, if the time horizon is long enough, intergenerational Pareto is violated.

Theorem 9. *Suppose each generation-t individual i's discounting utility function has an instantaneous utility function u_i and a discount function δ_i such that (A2) and (A3) hold and $(u_i)_{i\in N}$ is linearly independent. Let the planner's instantaneous utility function u be an arbitrary strict convex combination of $(u_i)_{i\in N}$. Then,*

1. *for each $\delta > \max_i \delta_i^*$, the planner is intergenerationally Pareto and strongly non-dictatorial;*

2. *for each $\delta < \max_i \delta_i^*$, there exists some $T^* > 0$ such that if $T \geq T^*$, the planner is not intergenerationally Pareto.*

Again, the cutoff is independent of the planner's risk attitude. Theorem 9 assumes (A2) and (A3). If we replace (A3) with (A1), as in Theorem 6, the only change in the statement of Theorem 9 will be that instead of one cutoff, we will have two cutoffs, as in Theorem 6.

The proof of this theorem is similar to Proposition II.3, with some new elements taken from the proof of Theorem 6. Similar to Theorem 6, the second part of Theorem 9 requires that the time horizon be long enough.

II.5.4 Transition of the Cutoff

Let us further illustrate how the cutoff changes from the least patient individual's long-run discount factor to the most patient individual's. An individual's instantaneous utility function describes his risk attitude. Let Θ be some positive integer between 1 and N. Suppose there is a linearly independent Θ-tuple of instantaneous utility functions $(u^\theta)_{\theta=1}^{\Theta}$ representing Θ generic types of risk attitude. Assume that individual i's instantaneous utility function $u_i \in \{u^\theta\}_{\theta=1}^{\Theta}$, and there is no redundant type in $\{u^\theta\}_{\theta=1}^{\Theta}$; that is, for each type u^θ, at least one individual's instantaneous utility function is equal to u^θ. If $\Theta = 1$, we are in the case of Theorem 6. If $\Theta = N$, we are in the case of Theorem 9. Define $\delta_\theta^* := \min_{k\in\{i\in N: u_i = u^\theta\}} \delta_k^*$; that is, for each θ, let δ_θ^* be the least patient individual's long-run discount factor whose type is u^θ. Define

$$\delta_{\text{maxmin}}^* := \max_\theta \delta_\theta^*.$$

Theorem 10. *Suppose for some linearly independent Θ-tuple of instantaneous utility functions $(u^\theta)_{\theta=1}^{\Theta}$, each generation-t individual i's discounting utility function has an instantaneous utility function $u_i \in \{u^\theta\}_{\theta=1}^{\Theta}$ and a discount function δ_i such that (A2) and (A3) hold*

and $\{u_i\}_{i\in N} = \{u^\theta\}_{\theta=1}^{\Theta}$. *Let the planner's instantaneous utility function* u *be an arbitrary strict convex combination of* $(u_i)_{i\in N}$. *Then,*

1. *for each* $\delta > \delta^*_{maxmin}$, *the planner is intergenerationally Pareto and strongly non-dictatorial;*
2. *for each* $\delta < \delta^*_{maxmin}$, *there exists some* $T^* > 0$ *such that if* $T \geq T^*$, *the planner is not intergenerationally Pareto.*

Note that as Θ increases from 1 to N, the cutoff may not increase monotonically. The idea of the proof of the theorem is as follows. For each type of risk attitude u^θ, we can apply Theorem 6 to show that the cutoff for the social discount factor implied by aggregating type-u^θ individuals is δ^*_θ. When aggregating across types, we apply Theorem 9 to show that the maximal δ^*_θ is the cutoff for the social discount factor.

II.6 Conclusion

The value of a policy or a public project that affects generations of individuals often crucially depends on which social discount rate is used for the evaluation. However, there is no consensus on which social discount rate is the right one to use. This paper considers a few important and widely used assumptions in economics, and characterizes the set of social discount rates that are compatible with those assumptions. The key assumptions are (i) individuals discount future consumption heterogeneously, (ii) the planner has an exponential discounting expected utility function, (iii) the planner is intergenerationally Pareto, which means that if all individuals from the current and future generations agree that one consumption sequence is better than another, the planner must agree, and (iv) the planner never completely ignores any individual's preference.

We show that for a generic set of individual instantaneous utility functions, social discounting should be more patient than the most patient individual's long-run discounting, as long as the time horizon is long enough, independent of the planner's instantaneous utility function. Therefore, using a near-zero social discount rate is justifiable in our framework.

II.7 Proof

II.7.1 Proof of Proposition II.1

Proof. **If Part** If there exists some $i \in N$ such that $U_t = U_{i,t}$ for any $t \in T$, the planner only takes individual i into account in period t. The corresponding weights in period t are $\omega_i = 1$ and $\omega_j = 0$ for all $j \neq i$. According to Lemma 3, the planner's preference $(\succsim_t)_{t\in T}$ is current-generation Pareto.

Only-If Part Suppose the planner is current-generation Pareto. We only prove the only-if part for the first period. According to Lemma 3, there exists an N-tuple of nonnegative weights $(\omega_i)_{i\in N}$, such that

$$\sum_{i=1}^{N} \omega_i \sum_{\tau=1}^{T} \delta_i^{\tau-1} u_i(p_\tau) = \sum_{\tau=1}^{T} \delta^{\tau-1} u(p_\tau);$$

that is, for $\tau = 1, \ldots, T-1$,

$$\sum_{i=1}^{N} \omega_i \delta_i^{\tau-1} u_i(p_\tau) = \delta^{\tau-1} u(p_\tau).$$

Let $\tau = 1$, 2, and 3. We have

$$\begin{cases} \sum_{i=1}^{N} \omega_i u_i(p) = u(p), \\ \sum_{i=1}^{N} \omega_i \delta_i u_i(p) = \delta u(p), \\ \sum_{i=1}^{N} \omega_i \delta_i^2 u_i(p) = \delta^2 u(p), \end{cases}$$

for any $p \in \Delta(X)$. Let $p = x^*$. The first equation shows that $\sum_{i\in N} \omega_i = 1$. Combining the second and the third equations above,

$$\left(\sum_{i=1}^{N} \omega_i \delta_i \right)^2 = \sum_{i=1}^{N} \omega_i \delta^2. \tag{II.10}$$

Since $\sum_{i\in N} \omega_i = 1$ and δ_i's are distinct, by Jensen's inequality, equation (II.10) holds if and only if there is some $i \in N$ such that $\omega_i = 1$ ($\omega_j = 0$ for any $j \neq i$). Thus, $U_1 = U_{i,1}$. Since the planner's instantaneous utility function and the social discount factor do not change over time, $U_t = U_{i,t}$ for any $t \in T$. □

II.7.2 Proof of Proposition II.2

Proof. The following lemma will be useful in proving Proposition II.2.

Lemma 11. *Given a positive N-tuple $(\delta_i)_{i\in N}$, for any $t \in T$, there exists a finite sequence of positive numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that*

$$\sum_{i=1}^{N}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i^{\tau-s} = \delta^{\tau-t} \tag{II.11}$$

for any $\tau \geq t$ if and only if $\delta > \min_i \delta_i$.

Proof. **If Part** Without loss of generality, we assume that $\delta_1 = \min_i \delta_i$. First, we fix all the weights other than individual 1's. Let $\omega_t(i,s) = \epsilon_t(s) > 0$ for any $i \geq 2$, $t \geq 1$, and $s \geq t$. The remaining part is to find $(\omega_t(1,s))_{t\in T, s\geq t}$ such that

1. equation (II.11) holds;
2. $\omega_t(1,s) > 0$, for any $t \geq 1$ and $s \geq t$.

Construct $(\omega_t(1,s))_{t\in T,s\geq t}$ by the following recursive formula:

$$\omega_t(1,s) = \begin{cases} 1 - \sum_{i=2}^{N}\omega_t(i,s), & \text{if } s = t, \\ \delta^{s-t} - \sum_{i=1}^{N}\omega_t(i,t)\delta_i^{s-t} - \cdots - \sum_{i=1}^{N}\omega_t(i,s-1)\delta_i - \sum_{i=2}^{N}\omega_t(i,s), & \text{if } s > t. \end{cases} \tag{II.12}$$

It can be verified that (II.12) ensures that equation (II.11) holds for any $t \in T$ and $\tau \geq t$. The remaining part is to show that $(\omega_{1,t}(s))_{t\in T,s\geq t}$ derived from (II.12) are strictly greater than zero, if $(\epsilon_t(s))_{t\in T,s\geq t}$ are small enough. We prove it in two steps.

Step 1 Setting $\epsilon_t(s) = 0$, the recursive formula (II.12) becomes

$$\omega_t(1,s) = \begin{cases} 1, & \text{if } s = t, \\ \delta^{s-t-1}(\delta - \delta_1), & \text{if } s > t, \end{cases}$$

for each $t \in T$. This can be proved by induction. Since $\delta > \delta_1$, we have $\omega_t(1,s) > 0$.

Step 2 Plugging $\epsilon_t(s)$ into formula (II.12), we have,

$$\begin{cases} \omega_t(1,t) = 1-(N-1)\epsilon_t(t), \\ \omega_t(1,t+1) = \delta - \delta_1 - \left[\sum_{i=2}^{N}(\delta_i-\delta_1)\right]\epsilon_t(t) - (N-1)\epsilon_t(t+1), \\ \omega_t(1,t+2) = \delta(\delta-\delta_1) - \left[\sum_{i=2}^{N}\delta_i(\delta_i-\delta_1)\right]\epsilon_t(t) - \left[\sum_{i=2}^{N}(\delta_i-\delta_1)\right]\epsilon_t(t+1) - (N-1)\epsilon_t(t+2), \\ \vdots \end{cases}$$

Then, we know that $\omega_t(1,s) = F_t^{(s)}(\epsilon_t(t),\ldots,\epsilon_t(s)|\delta,\delta_1,\ldots,\delta_n)$, in which $F_t^{(s)}$ is an affine (and hence continuous) function of $\epsilon_t(t),\ldots,\epsilon_t(s)$. Since $F_t^{(s)}$ is continuous, the weights $\omega_t(1,s)$'s are strictly greater than zero, if $\epsilon_t(s)$'s are small enough.

Only-If Part For any $t<T$, let $\tau = t, t+1$ in (II.11). We have

$$\begin{cases} \sum_{i=1}^{N}\omega_t(i,t) = 1, \\ \sum_{i=1}^{N}\omega_t(i,t)\delta_i + \sum_{i=1}^{N}\omega_t(i,t+1) = \delta. \end{cases}$$

Combining the above two equations,

$$\sum_{i=1}^{N}\omega_t(i,t)\delta = \sum_{i=1}^{N}\omega_t(i,t)\delta_i + \sum_{i=1}^{N}\omega_t(i,t+1).$$

Rearranging the above equation, we have

$$\delta = \frac{\sum_{i=1}^{N}\omega_t(i,t)\delta_i + \sum_{i=1}^{N}\omega_t(i,t+1)}{\sum_{i=1}^{N}\omega_t(i,t)} > \frac{\sum_{i=1}^{N}\omega_t(i,t)\delta_i}{\sum_{i=1}^{N}\omega_t(i,t)} > \frac{\sum_{i=1}^{N}\omega_t(i,t)\min_i\delta_i}{\sum_{i=1}^{N}\omega_t(i,t)} = \min_{i\in N}\delta_i.$$

□

Now we prove Proposition II.2.

If Part Taking the weights from the if part of Lemma 11, since the planner's instantaneous utility function u is identical to individual instantaneous utility function u, we immediately know that the planner has the desired EDU function, because

$$\begin{aligned} U_t(\mathbf{p}) &= \sum_{s=t}^{T}\sum_{i=1}^{N}\omega_t(i,s)U_{i,s}(\mathbf{p}) = \sum_{s=t}^{T}\sum_{i=1}^{N}\omega_t(i,s)\sum_{\tau=s}^{T}\delta_i^{\tau-s}u(p_\tau) \\ &= \sum_{\tau=t}^{T}\sum_{s=t}^{\tau}\sum_{i=1}^{N}\omega_t(i,s)\delta_i^{\tau-s}u(p_\tau) = \sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau). \end{aligned}$$

Only-If Part Suppose the planner's preference is intergenerationally Pareto and strongly non-dictatorial. For each $t \in T$, there exists a finite sequence of positive numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that

$$U_t(\mathbf{p}) = \sum_{s=t}^{T}\sum_{i=1}^{N} \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=t}^{T}\sum_{s=t}^{\tau}\sum_{i=1}^{N} \omega_t(i,s)\delta_i^{\tau-s} u(p_\tau).$$

Then, for any $t \in T$ and $\tau \geq t$, the following equality holds:

$$\sum_{s=t}^{\tau}\sum_{i=1}^{N} \omega_t(i,s)\delta_i^{\tau-s} u(p_\tau) = \delta^{\tau-t} u(p_\tau).$$

Let $p_\tau = x^*$. Lemma 11 implies that $\delta > \min_i \delta_i$. □

II.7.3 One-Individual Intergenerational Aggregation

We prove the following lemma for intergenerational aggregation when each generation only has one individual.

Lemma 12. *Assume that $N = \{i\}$. Suppose each generation-t individual i's discounting utility function has an instantaneous utility function u and a discount function δ_i such that (A1) and (A2) hold. For any $\delta > \hat{\delta}_i := \max_{\tau\in\{0,\ldots,T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$, the planner is intergenerationally Pareto and strongly non-dictatorial.*

Proof. We want to show that for any $\delta > \hat{\delta}_i$ and $t \in T$, there exists a finite sequence of positive numbers $(\omega_t(i,s))_{s\geq t}$ such that

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u(p_\tau) = \sum_{s=t}^{T} \omega_t(i,s) U_{i,s}(\mathbf{p}).$$

Given any $\delta > \hat{\delta}_i$, for each $t \in T$, we can construct $(\omega_t(i,s))_{s\geq t}$ according to the following formula:

$$\omega_t(i,s) = \begin{cases} 1, & \text{if } s = t, \\ \delta^{s-t-1}\left(\delta - \hat{\delta}_i\right) + \sum_{\tau=t}^{s-1}\left[\hat{\delta}_i\delta_i(s-1-\tau) - \delta_i(s-\tau)\right]\omega_t(i,\tau), & \text{if } s > t. \end{cases} \tag{II.13}$$

Note that by assuming $\delta > \hat{\delta}_i$, for $s > t$, the first term of $\omega_t(i,s)$ is strictly greater than 0. According to the definition of $\hat{\delta}_i$, the second term of $\omega_t(i,s)$ is greater than 0. Hence,

$\omega_t(i,s) > 0$ for any $s \geq t$. Then,

$$U_t(\mathbf{p}) = \sum_{s=t}^{T} \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{s=t}^{T} \omega_t(i,s) \left[\sum_{\tau=s}^{T} \delta_i(\tau - s) u(p_\tau) \right] = \sum_{\tau=t}^{T} \left[\sum_{s=t}^{\tau} \delta_i(\tau - s) \omega_t(i,s) \right] u(p_\tau).$$

We want to prove that $U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u(p_\tau)$ by induction. Consider $\sum_{s=t}^{\tau} \delta_i(\tau - s)\omega_t(i,s)$. When $\tau = t$, $\sum_{s=t}^{\tau} \delta_i(\tau - s)\omega_t(i,s) = \omega_t(i,t) = 1 = \delta^0$. Suppose for some $\tau \geq t$, we have proven that $\sum_{s=t}^{\tau} \delta_i(\tau - s)\omega_t(i,s) = \delta^{\tau-t}$. We want to prove that for $\tau + 1$,

$$\sum_{s=t}^{\tau+1} \delta_i(\tau + 1 - s)\omega_t(i,s) = \delta^{\tau-t+1}. \tag{II.14}$$

To prove (II.14), we only need to notice that according to (II.13),

$$\begin{aligned}
&\sum_{s=t}^{\tau+1} \delta_i(\tau + 1 - s)\omega_t(i,s) = \omega_t(i,\tau+1) + \sum_{s=t}^{\tau} \delta_i(\tau + 1 - s)\omega_t(i,s) \\
= \;& \omega_t(i,\tau+1) + \hat{\delta}_i \left[\delta^{\tau-t} + \sum_{s=t}^{\tau} \frac{\delta_i(\tau+1-s)}{\hat{\delta}_i} \omega_t(i,s) - \delta^{\tau-t} \right] \\
= \;& \omega_t(i,\tau+1) + \hat{\delta}_i \left[\delta^{\tau-t} + \sum_{s=t}^{\tau} \frac{\delta_i(\tau+1-s)}{\hat{\delta}_i} \omega_t(i,s) - \sum_{s=t}^{\tau} \delta_i(\tau - s)\omega_t(i,s) \right] \\
= \;& \omega_t(i,\tau+1) + \hat{\delta}_i \delta^{\tau-t} + \sum_{s=t}^{\tau} \left[\delta_i(\tau+1-s) - \hat{\delta}_i \delta_i(\tau - s) \right] \omega_t(i,s) = \delta^{\tau-t+1}.
\end{aligned}$$

By induction, we know that $\sum_{s=t}^{\tau} \delta_i(\tau - s)\omega_t(i,s) = \delta^{\tau-t}$ for any $\tau \geq t$, and hence $U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u_i(p_\tau)$. □

II.7.4 Proof of Proposition II.3

Proof. **If Part** Lemma 12 states that the planner can find a sequence of positive numbers $(\tilde{\omega}_t(i,s))_{t\in T, i\in N, s\geq t}$ such that

$$\sum_{s=1}^{T} \tilde{\omega}_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=1}^{T} \delta^{\tau-1} u_i(p_\tau),$$

for any $i \in N$ and consumption sequence $\mathbf{p}$, because $\delta > \max_i \hat{\delta}_i = \max_i \delta_i$. Now, since the planner's instantaneous utility function $u = \sum_{i\in N} \lambda_i u_i$, we only need to let $\omega(i,t) = \lambda_i \tilde{\omega}(i,t)$.

Then,

$$\sum_{i=1}^{N}\sum_{s=1}^{T}\omega_t(i,s)U_{i,s}(\mathbf{p})=\sum_{i=1}^{N}\sum_{s=1}^{T}\lambda_i\tilde{\omega}_t(i,s)U_{i,s}(\mathbf{p})=\sum_{i=1}^{N}\lambda_i\left[\sum_{\tau=1}^{T}\delta^{\tau-1}u_i(p_\tau)\right]=\sum_{\tau=1}^{T}\delta^{\tau-1}u(p_\tau).$$

Only-If Part Note that when $(u_i)_{i\in N}$ is linearly independent and u is in the interior of $\text{co}(\{u_i\}_{i\in N})$, there is a unique way to write u as a strict convex combination of $(u_i)_{i\in N}$. Suppose $\sum_{i\in N}\lambda_i u_i = u$, in which $\lambda_i > 0$ and $\sum_{i\in N}\lambda_i = 1$. We only need to consider the period-1 planner. The planner's discounting utility function satisfies

$$U_1(\mathbf{p})=\sum_{s=1}^{T}\sum_{i=1}^{N}\omega_1(i,s)U_{i,s}(\mathbf{p})=\sum_{s=1}^{T}\sum_{i=1}^{N}\omega_1(i,s)\sum_{\tau=s}^{T}\delta_i^{\tau-s}u_i(p_\tau), \tag{II.15}$$

in which $\omega_1(i,s) > 0$ is the weight the period-1 planner assigns to the generation-s individual i. Since $\mathbf{p}$ is arbitrary, this implies that the planner's instantaneous utility function for period-1 consumption satisfies

$$u(p_1)=\sum_{i=1}^{N}\omega_1(i,1)u_i(p_1)$$

for any p_1. Because $u=\sum_{i\in N}\lambda_i u_i$ and $(u_i)_{i\in N}$ is linearly independent,

$$\omega_1(i,1)=\lambda_i \tag{II.16}$$

must hold for any $i \in N$. Similarly, equation (II.15) implies that for period-2 consumption,

$$\delta u(p_2)=\sum_{i=1}^{N}[\omega_1(i,1)\delta_i+\omega_1(i,2)]u_i(p_2)$$

for any p_2. Since instantaneous utility functions do not depend on time, the unique way to write u as a strict convex combination of $(u_i)_{i\in N}$ does not change; that is, $\delta u(p_2) = \delta\sum_{i\in N}\lambda_i u_i(p_2)$ for any p_2. Then, for any $i \in N$,

$$\lambda_i\delta=\omega_1(i,1)\delta_i+\omega_1(i,2). \tag{II.17}$$

Combining equations (II.16) and (II.17) gives us

$$\delta=\delta_i+\frac{\omega_1(i,2)}{\omega_1(i,1)}>\delta_i,$$

for any $i \in N$. The last strict inequality follows from the fact that $\omega_1(i,s) > 0$. Therefore, $\delta > \max_i \delta_i$. □

II.7.5 Proof of Theorem 6

Proof. **Part I** We prove the first part in three steps.

In the first step, the period-t planner does $T-t+1$ times aggregations for each individual i. The σ^{th} aggregation aggregates individual i from generation σ to generation T (σ starts from t and ends at T), and it assigns weight $\tilde{\omega}_{t,\sigma}(i,s)$ to generation-s individual i for $s \geq \sigma$.

By Lemma 12, we know that for each $t \in T$, $i \in N$ and $\sigma \geq t$, the planner can find a sequence of positive weights $(\tilde{\omega}_{t,\sigma}(i,s))_{s\geq\sigma}$ such that

$$\sum_{s=\sigma}^{T} \tilde{\omega}_{t,\sigma}(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=\sigma}^{T} \delta^{\tau-\sigma} u(p_\tau),$$

for any $\delta > \hat{\delta}_i$. The σ^{th} aggregation, as if, gives the planner an exponential discounting generation-σ individual i with a discount factor slightly higher than $\hat{\delta}_i$.

In the second step, the period-t planner collects all of the weights assigned to generation-s individual i for all $T-t+1$ times step-one aggregations. Then the weight assigned to generation-s individual i by the period-t planner is

$$\tilde{\omega}_t(i,s) = \sum_{\sigma=s}^{T} \tilde{\omega}_{t,\sigma}(i,s).$$

Essentially, in each period t, the step-two aggregation under weights $(\tilde{\omega}_t(i,s))_{t\in T, s\geq t}^{i\in N}$ gives the planner N exponential discounting individuals from the t^{th} generation to the T^{th} generation, and each individual has a discount factor slightly higher than $\hat{\delta}_i$.

Lastly, by Proposition II.2, the planner can aggregate N exponential discounting individuals one more time, and obtain an EDU function with any social discount factor greater than $\min_i \hat{\delta}_i$.

Part II Define $\tilde{\delta}_i := \lim_{\tau\to\infty} \sqrt[\tau]{\delta_i(\tau)}$. We assume that $\tilde{\delta}_1$ is the unique minimum of $\tilde{\delta}_1, \ldots, \tilde{\delta}_N$. The proof can easily be extended to the case with multiple minima. We prove it by contradiction. Suppose the planner is intergenerationally Pareto. For each $t \in T$, there exists a finite sequence of nonnegative numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that the following equality holds:

$$\sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s) \delta_i(\tau - s) u(p_\tau) = \delta^{\tau-t} u(p_\tau) \tag{II.18}$$

for any $t \in T$ and $\tau \geq t$.

By letting $\tau = t$, equation (II.18) shows that $\sum_{i \in N} \omega_t(i,t) = 1$ for any $t \in T$. Then,

$$\delta^{\tau - t} = \frac{\sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s) \delta_i(\tau - s)}{\sum_{i=1}^{N} \omega_t(i,t)} \geq \frac{\sum_{i=1}^{N} \omega_t(i,t) \delta_i(\tau - t)}{\sum_{i=1}^{N} \omega_t(i,t)}. \tag{II.19}$$

Since $\tilde{\delta}_1 = \min_i \tilde{\delta}_i$, there exists $T_1 > 0$ such that for each $\tau > T_1$, $\delta_1(\tau - t) = \min_i \delta_i(\tau - t)$. Hence, (II.19) becomes

$$\delta^{\tau - t} \geq \frac{\sum_{i=1}^{N} \omega_{i,t}(t) \delta_1(\tau - t)}{\sum_{i=1}^{N} \omega_{i,t}(t)} = \delta_1(\tau - t). \tag{II.20}$$

According to our assumptions, $\delta < \tilde{\delta}_1$. Then, there exists $T_2 > 0$ such that for each $\tau > T_2$,

$$\delta^{\tau - t} < \delta_1(\tau - t). \tag{II.21}$$

Let $T^* = \max\{T_1, T_2\}$. Then, (II.20) and (II.21) contradict each other. □

II.7.6 Proof of Theorem 9

Proof. **Part I** We prove this theorem in two steps. First, we again consider the special case in which there is only one individual i to be aggregated across generations. Since the individual relative discount factor is nondecreasing, $\delta_i^* \geq \hat{\delta}_i := \max_{\tau \in \{0, \ldots, T-2\}} \frac{\delta_i(\tau+1)}{\delta_i(\tau)}$. By Lemma 12, because the social discount factor $\delta > \max_i \delta_i^* \geq \delta_i^*$, for any $i \in N$ and $t \in T$, we can find some positive $(\omega_t(i,s))_{s \geq t}$ such that

$$\sum_{s=t}^{T} \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau - t} u_i(p_\tau);$$

that is, we can aggregate each individual's utility functions across generations into an EDU function with discount factor δ.

Consider any N-tuple of positive numbers $(\lambda_i)_{i \in N}$ such that $\sum_{i \in N} \lambda_i = 1$. Together with the weights $(\omega_t(i,s))_{t \in T, i \in N, s \geq t}$ we have found above, let the planner's utility function

satisfy

$$U_t(\mathbf{p}) = \sum_{i=1}^{N}\sum_{s=t}^{T}\lambda_i\omega_t(i,s)U_{i,s}(\mathbf{p}) = \sum_{i=1}^{N}\sum_{\tau=t}^{T}\delta^{\tau-t}\lambda_i u_i(p_\tau)$$
$$= \sum_{\tau=t}^{T}\delta^{\tau-t}\sum_{i=1}^{N}\lambda_i u_i(p_\tau) = \sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau),$$

in which $u = \sum_{i\in N}\lambda_i u_i$ is an arbitrary strict convex combination of $(u_i)_{i\in N}$.

Part II We prove it by contradiction. Suppose there exists an intergenerationally Pareto planner with the social discount factor $\delta < \max_i \delta_i^*$. By intergenerational Pareto, for any $t \in T$, there exists nonnegative numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that the following equality holds for any $t \in T$:

$$\sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau) = \sum_{s=t}^{T}\sum_{i=1}^{N}\omega_t(i,s)\sum_{\tau=s}^{T}\delta_i(\tau-s)u_i(p_\tau) = \sum_{\tau=t}^{T}\sum_{i=1}^{N}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)u_i(p_\tau).$$

Since $\mathbf{p}$ is arbitrary, the equation above implies that for any $t \in T$ and $\tau \geq t$,

$$\sum_{i=1}^{N}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)u_i(p_\tau) = \delta^{\tau-t}u(p_\tau). \tag{II.22}$$

Recall that u is a strict convex combination of $(u_i)_{i\in N}$ and $(u_i)_{i\in N}$ is linearly independent. There is a unique way to write u as a convex combination of $(u_i)_{i\in N}$. Moreover, when $\tau = t$, equation (II.22) becomes

$$\sum_{i=1}^{N}\omega_t(i,t)u_i(p_t) = u(p_t) \tag{II.23}$$

for any $t \in T$. Thus, $\omega_t(i,t) > 0$ for any $i \in N$ and $t \in T$. Combining equations (II.22) and (II.23), we have

$$\delta^{\tau-t}\sum_{i=1}^{N}\omega_t(i,t)u_i(p_\tau) = \sum_{i=1}^{N}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)u_i(p_\tau).$$

Since $(u_i)_{i\in N}$ is linearly independent, for any $i \in N$, $t \in T$, and $\tau \geq t$, the above equation implies

$$\omega_t(i,t)\delta^{\tau-t} = \sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s),$$

which in turn implies

$$\begin{aligned}\delta^{\tau-t} &= \frac{\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)}{\omega_t(i,t)} = \frac{\omega_t(i,t)\delta_i(\tau-t)+\sum_{s=t+1}^{\tau}\omega_t(i,s)\delta_i(\tau-s)}{\omega_t(i,t)} \\ &\geq \frac{\omega_t(i,t)\delta_i(\tau-t)}{\omega_t(i,t)} = \delta_i(\tau-t);\end{aligned}$$

that is,

$$\delta \geq \sqrt[\hat{\tau}]{\delta_i(\hat{\tau})} \tag{II.24}$$

for any $1 \leq \hat{\tau} < T$.

Without loss of generality, we assume δ_N^* is a maximum of $\{\delta_i^*\}_{i\in N}$. Since $\delta < \delta_N^* = \lim_{\tau\to\infty}\sqrt[\tau]{\delta_N(\tau)}$, there exists T^* such that for any $T \geq T^*$, $\delta < \sqrt[T-1]{\delta_N(T-1)}$, which contradicts (II.24). □

II.7.7 Proof of Theorem 10

Proof. **Part I** We prove Part I in two steps. First, we aggregate individuals who share the same u^θ. For each $\theta \in \Theta$, $I^\theta := \{i \in N : u_i = u^\theta\}$ is called a "family," which is the set of i's whose instantaneous utility functions are u^θ. By Corollary 8, we know that for each θ and each $\delta > \min_{i\in I_\theta}\delta_i^*$, there exists a sequence of weights $(\omega_t(i,s))_{t\in T, i\in I_\theta, s\geq t}$ such that

$$U_t^\theta(\mathbf{p}) = \sum_{\tau=t}^{T}\delta^{\tau-t}u^\theta(p_\tau) = \sum_{s=t}^{T}\sum_{i\in I_\theta}\omega_t(i,s)U_{i,s}(\mathbf{p}).$$

for each $t \in T$. Now, we have $|\Theta|$ exponential discounting expected utility functions U_t^θ's with linearly independent instantaneous utility functions u^θ's.

Next, we apply Proposition II.3 to aggregate U_t^θ's. It follows immediately that if $\delta > \max_{\theta\in\Theta}\min_{i\in I_\theta}\delta_i^*$, the planner is intergenerationally Pareto and strongly non-dictatorial.

Part II We prove its contrapositive. Suppose the planner is intergenerationally Pareto. Then, for each $t \in T$, there exists a finite sequence of positive numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau) = \sum_{s=t}^{T}\sum_{\theta\in\Theta}\sum_{i\in I_\theta}\omega_t(i,s)\sum_{\tau=s}^{T}\delta_i(\tau-s)u_i(p_\tau),$$

and hence

$$\delta^{\tau-t}u(p_\tau) = \sum_{\theta\in\Theta}\sum_{i\in I_\theta}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)u^\theta(p_\tau) \tag{II.25}$$

for any $t \in T$ and $\tau \geq t$.

By letting $\tau = t$ in equation (II.25), we have

$$u(p_t) = \sum_{\theta\in\Theta}\sum_{i\in I_\theta} \omega_t(i,t)u^\theta(p_t). \tag{II.26}$$

Recall that u is a strict convex combination of $(u_i)_{i\in N}$. Equation (II.26) shows that $\sum_{i\in I_\theta}\omega_t(i,t) > 0$ for each θ. Combining equations (II.25) and (II.26), we have

$$\sum_{\theta\in\Theta}\sum_{i\in I_\theta}\delta^{\tau-t}\omega_t(i,t)u^\theta(p_\tau) = \sum_{\theta\in\Theta}\sum_{i\in I_\theta}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)u^\theta(p_\tau).$$

Since $(u^\theta)_{i=1}^{\Theta}$ is linearly independent, the above equation is equivalent to

$$\sum_{i\in I_\theta}\delta^{\tau-t}\omega_t(i,t) = \sum_{i\in I_\theta}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)$$

for any $\theta\in\Theta$. Rearranging the above equation, we obtain

$$\delta^{\tau-t} = \frac{\sum_{i\in I_\theta}\sum_{s=t}^{\tau}\omega_t(i,s)\delta_i(\tau-s)}{\sum_{i\in I_\theta}\omega_t(i,t)} > \frac{\sum_{i\in I_\theta}\omega_t(i,t)\delta_i(\tau-t)}{\sum_{i\in I_\theta}\omega_t(i,t)}. \tag{II.27}$$

Letting τ go to infinity, it is easy to see that (II.27) becomes $\delta \geq \min_{i\in I_\theta}\delta_i^*$ for $\forall\theta\in\Theta$. Hence, $\delta \geq \max_{\theta\in\Theta}\min_{i\in I_\theta}\delta_i^* = \delta^*_{\text{maxmin}}$. □

II.7.8 Infinite Time Horizon

Our findings can be extended to the case with $T = +\infty$. When $T = +\infty$, we require that individual discount factors $(\delta_i(\tau))_{\tau=0}^{\infty}$ be an absolutely summable sequence (in ℓ^1) and $\max_i\delta_i^* < 1$, and that the social discount factor $\delta < 1$. The result below will show that even when individual instantaneous utility functions are identical, the cutoff for the social discount factor will jump from $\min_i\delta_i^*$ to $\max_i\delta_i^*$ when T becomes infinite. We first define intergenerational utilitarianism.

Definition 13. The planner is *intergenerationally utilitarian* if in each period $t\in T$, there exists a sequence of nonnegative numbers $(\omega_t(i,s))_{i\in N,s\geq t}$ such that $0 < \sum_{i=1}^{N}\sum_{s=t}^{T}\omega_t(i,s) < \infty$, and

$$U_t = \sum_{i=1}^{N}\sum_{s=t}^{T}\omega_t(i,s)U_{i,s}.$$

Below, we will assume intergenerational utilitarianism rather than intergenerational

Pareto, because the equivalence between intergenerational utilitarianism and intergenerational Pareto for countably infinitely many individuals is not yet established.

Proposition II.4. *Suppose $T = +\infty$, and each generation-t individual i's discounting utility function has an instantaneous utility function u_i and a discount function δ_i such that (A2) and (A3) hold. Let the planner's instantaneous utility function u be an arbitrary strict convex combination of $(u_i)_{i\in N}$. Then,*

1. *for each $\max_i \delta_i^* < \delta < 1$, the planner is intergenerationally utilitarian and strongly non-dictatorial;*

2. *for each $\delta < \max_i \delta_i^*$, the planner is not simultaneously intergenerationally utilitarian and strongly non-dictatorial.*

Proof. **Part I** Since u is a strict convex combination of $(u_i)_{i\in N}$, suppose $u = \sum_i \lambda_i u_i$ for some $\lambda_1, \ldots, \lambda_N > 0$ such that $\sum_i \lambda_i = 1$. For each $i \in N$ and each $t \in T$, we want to construct a sequence of positive and absolutely summable numbers $(\omega_t(i,s))_{s=t}^{\infty}$ such that

$$\sum_{s=t}^{\infty} \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_i(p_\tau).$$

If this can be done, then in period t, let $\lambda_i \omega_t(i,s)$ be the planner's utilitarian weight for the generation-s individual i, in which case

$$\sum_{i=1}^{N} \sum_{s=t}^{\infty} \lambda_i \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u(p_\tau) = U_t(\mathbf{p}),$$

which means that the planner is intergenerationally utilitarian and strongly non-dictatorial.

Next, we show that the following recursive definition of $(\omega_t(i,s))_{s=t}^{\infty}$ works: For each $s \geq t$,

$$\omega_t(i,s) = \begin{cases} 1, & \text{if } s = t, \\ \sum_{\sigma=t}^{s-1} [\delta \cdot \delta_i(s-\sigma) - \delta_i(s-\sigma+1)] \omega_t(i,\sigma), & \text{if } s > t. \end{cases} \tag{II.28}$$

First, it can be verified that each $\omega_t(i,s)$ is positive, because $\delta > \max_i \delta_i^*$ and the individual relative discount factor is nondecreasing. Second, it can be verified inductively that for any finite τ,

$$\sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s) \delta_i(\tau - s) u_i(p_\tau) = \delta^{\tau-t} u(p_\tau)$$

for any $p_\tau \in \Delta(X)$. These two steps are similar to the steps in the proof of Lemma 12. Thus, we only have to show that $(\omega_t(i,s))_{s=t}^{\infty}$ is summable. Clearly, $\sum_{s=t}^{n} \omega_t(i,s)$ is nondecreasing

in n. If we can show that $\sum_{s=t}^{n} \omega_t(i,s)$ is bounded above and the bound is a constant, this part of the theorem is proven.

Sum up both sides of equation (II.28) from $s = t$ to n. We obtain that

$$1 = \sum_{s=t}^{n-1} \left((1-\delta) \sum_{\tau=0}^{n-1-s} \delta_i(\tau) + \delta_i(n-s) \right) \omega_t(i,s) + \omega_t(i,n).$$

Because $\sum_{\tau=0}^{n-1-s} \delta_i(\tau) > 1$ and $\delta_i(n-s) > 0$, $(1-\delta)\sum_{\tau=0}^{n-1-s} \delta_i(\tau) + \delta_i(n-s) > 1-\delta$, which implies that

$$1 > \sum_{s=t}^{n-1} (1-\delta)\omega_t(i,s) + \omega_t(i,n) > (1-\delta) \sum_{s=t}^{n-1} \omega_t(i,s).$$

Therefore, $\sum_{s=t}^{n-1} \omega_t(i,s)$ is bounded above by $1/(1-\delta)$ for any n.

Part II Assume that δ_N^* is the unique maximum of $\{\delta_i^*\}_{i \in N}$. The proof can easily be extended to the case with multiple maxima. We prove the contrapositive of this part. Suppose the planner is intergenerationally utilitarian and strongly non-dictatorial; that is, for each $t \in T$, there exists a sequence of positive numbers $(\omega_t(i,s))_{i \in N, s \geq t}$ such that $U_t = \sum_{i,s} \omega_t(i,s) U_{i,s}$. Hence, for any $t \in T$ and $\tau \geq t$,

$$\sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s)\delta_i(\tau - s)u_i(p_\tau) = \delta^{\tau - t} u(p_\tau).$$

Consider a consumption sequence that yields x^* in every period, $(x^*, x^*, \dots)$. Then, the equation above becomes

$$\sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s)\delta_i(\tau - s) = \delta^{\tau - t}.$$

Since u_i's and u are normalized, we know that for each t, $\sum_{i \in N} \omega_t(i,t) = 1$. Due to the strongly non-dictatorial property, in particular, $\omega_t(N,t) \in (0,1)$. Then,

$$\delta^{\tau - t} = \sum_{s=t}^{\tau} \sum_{i=1}^{N} \omega_t(i,s)\delta_i(\tau - s) > \omega_t(N,t)\delta_N(\tau - t).$$

Therefore, $\delta > \sqrt[\tau - t]{\omega_t(N,t)\delta_N(\tau - t)}$ for every τ implies that $\delta \geq \delta_N^*$. □

Proposition II.4 covers the case in which u_i's are identical. Thus, Proposition II.4 says that if $T = +\infty$, the cutoff for the social discount factor again jumps from $\min_i \delta_i^*$ to $\max_i \delta_i^*$, compared to Theorem 6/Corollary 8.

Note that the second part of Proposition II.4 is weaker than the second part of Theorem

6, Corollary 8, or Theorem 9. In Proposition II.4, if the social discount factor is lower than the highest individual long-run discount factor, the conclusion is that either intergenerational utilitarianism is violated or the planner has ignored some individual from some generation.

Nonetheless, there is still discontinuity between Proposition II.4 and Theorem 6/Corollary 8. In Theorem 6/Corollary 8, if the social discount factor is lower than the lowest individual long-run discount factor, we know that intergenerational Pareto is violated, which implies that at least one of the two properties, intergenerational utilitarianism or the strongly non-dictatorial property, is violated, as in Proposition II.4.

The intuition for this discontinuity is the following. For simplicity, suppose u_i's are the same. Fixing an arbitrarily large but finite T, the planner can always attach small enough utilitarian weights to individuals with high δ_i^*. In this way, the planner can keep her social discount factor low. However, if T is infinite, fixing any positive weights, as τ increases to infinity, $\delta_i(\tau)$ of the individual with the highest δ_i^* dominates all other individuals' discount factors regardless of his weight. Therefore, the social discount factor cannot be strictly less than $\max_i \delta_i^*$.

II.8 Supplemental Material

This supplement consists of four parts: (i) the robustness of findings in the main paper with respect to several main assumptions, (ii) an alternative interpretation of intergenerational Pareto and its implication on quasi-hyperbolic discounting, (iii) a result with forward and backward individual exponential discounting, and (iv) a discussion of the choice domain of the main chapter.

II.8.1 Discussion of the Main Assumptions

Our main findings are built upon three sets of assumptions: (i) the assumptions about individual preferences, (ii) intergenerational Pareto and strong non-dictatorship, and (iii) the assumptions about the planner's preference. In the first, we have assumed that a parent's discount function and instantaenous utility function are inherited by his offspring. This assumption may or may not be realistic. It is helpful to understand how our results depend on it. In the second, intergenerational Pareto only has bite when all individuals from the current and future generations agree. It is useful to understand to what extent intergenerational Pareto can be strengthened. In the third, we have required that the planner have an EDU function. This assumption imposes restrictions on how the planner can aggregate individual preferences. We examine what results still hold if we drop this assumption.

We first state a more general version of Lemma 3, which also follows from Harsanyi (1955) and Fishburn (1984) directly.

Lemma 14. *(Harsanyi (1955)) Suppose each generation-t individual i's utility function takes the following form:*

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_{i,t}(\tau - t) u_i(p_\tau, \tau),$$

and the planner's utility function in period t takes the following form:

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_t(\tau - t) u_t(p_\tau, \tau),$$

in which $\delta_{i,t}(\cdot)$ and $\delta_t(\cdot)$ are discount functions, and $u_i(\cdot, \tau)$ and $u_t(\cdot, \tau)$ are normalized instantaneous utility functions. The planner's preference $(\succsim_t)_{t\in T}$ is intergenerationally Pareto if and only if in each period $t \in T$, there exists a finite sequence of nonnegative numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that

$$U_t = \sum_{i=1}^{N} \sum_{s=t}^{T} \omega_t(i,s) U_{i,s}.$$

We stick to the assumptions about individuals' and the planner's utility functions in the main paper, unless stated otherwise.

Inheriting Discount Functions and Instantaneous Utility Functions from Parents

One maintained assumption about individual preferences is that each generation-t individual i's discount function δ_i and instantaneous utilty function u_i are independent of t. We show in this subsection that this assumption can be removed without changing our main findings. We analyze two cases below. In the first case, for any $i \in N$ and finite t, suppose generation-t individual i's discount function is $\delta_{i,t}$ and instantaneous utility function is u_i; that is, we still assume that individual instantaneous utility functions do not depend on time. Fixing each generation-t individual i's discounting utility function for any $i \in N$ and *any natural number* t, our result may require us to vary the time horizon T. The result below shows that we can establish a positive result that is similar to Theorem 9.

Theorem 15. *Suppose each generation-t individual i's discounting utility function has an instantaneous utility function u_i and a discount function $\delta_{i,t}$ such that (A2) and (A3) hold and $(u_i)_{i\in N}$ is linearly independent. Let the planner's instantaneous utility function u be an arbitrary strict convex combination of $(u_i)_{i\in N}$. Then,*

1. *for each* $\delta > \max_{i,t} \delta^*_{i,t}$*, the planner is intergenerationally Pareto and strongly non-dictatorial;*

2. *for each* δ *such that for some* i, t, $\delta < \delta^*_{i,t}$*, there exists some* $T^* > 0$ *such that if* $T \geq T^*$*, the planner is not intergenerationally Pareto.*

We will prove this theorem as a special case of Theorem 17 below. Theorem 15 shows that social discounting should still be more patient than the most patient individual's long-run discounting when individual discount functions may change across generations. Since generation-t individual i's discount function is now $\delta_{i,t}$ rather than δ_i, the cutoff for the social discount factor becomes $\max_{i,t} \delta^*_{i,t}$. The second part of the theorem can be understood as follows. Suppose the social discount factor δ is below some generation-t individual i's long-run discount factor. Then, as we increase T, this planner will eventually violate intergenerational Pareto.

One may wonder why we still assume that generation-t individual i's instantaneous utility function does not depend on t. Let us assume that generation-t individual i's instantaneous utility function is $u_{i,t}$. The example below shows that this assumption will lead to a trivial negative result that has nothing to do with discounting.

Example 16. Suppose $N = 1$. Let generation-1 individual's instantaneous utility function be u_1, which is linearly independent of generation-2 individual's instantaneous utility function u_2. Since the planner has an EDU function, her instantaneous utility function should never change. In the first period, the planner's instantaneous utility function for period-1 consumption can only be u_1, because only the generation-1 individual cares about period-1 consumption. The planner's instantaneous utility function for period-2 consumption, however, must depend on both u_1 and u_2 due to strong non-dictatorship, which means that the planner's instantaneous utility function for period-2 consumption must differ from u_1. Therefore, it is impossible to require that the planner be intergenerationally Pareto and strongly non-dictatorial.

As can be seen in the example above, it seems inevitable that the planner's instantaneous utility function should depend on time; that is, the planner's instantaneous utility function for period-τ consumption should depend on τ. Indeed, one way to restore the positive result is to allow the planner's instantaneous utility function to be $u(\cdot, \tau)$.

However, there is another way to restore the positive result, which is the second case we want to analyze. For any $i \in N$ and finite t, suppose generation-t individual i's discount function is $\delta_{i,t}$ and instantaneous utility function for period-τ consumption is $u_i(\cdot, \tau)$; that is, if the planner's instantaneous utility function for period-τ consumption has to depend on τ,

let us make the same assumption for individuals. Note that individual instantaneous utility functions now depend on time, but in a manner different from Example 16. The planner's discount function is again exponential.

These assumptions are particularly suitable in our setting. Recall that each individual only lives for one period, and he cares about future consumption based on altruism. Imagine that $u_i(\cdot, \tau)$ is generation-τ individual i's actual consumption utility—that is, the utility that generation-τ individual i derives by consuming rather than from altruism. Now, generation-t individual i's utility function is

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta_i(\tau - t) u_i(p_\tau, \tau),$$

which means that when the generation-t individual i altruistically cares about generation-τ individual i's consumption, he values the consumption in exactly the same way that generation-τ individual i will value it for himself.

Theorem 17. *Suppose each generation-t individual i's discounting utility function has instantaneous utility functions $(u_i(\cdot, \tau))_{\tau \geq t}$ and a discount function $\delta_{i,t}$ such that (A2) and (A3) hold and $(u_i(\cdot, \tau))_{i \in N}$ is linearly independent for each $\tau \in T$. Suppose for some positive $(\lambda_i)_{i \in N}$ such that $\sum_{i \in N} \lambda_i = 1$, the planner's $u(\cdot, \tau) = \sum_{i \in N} \lambda_i u_i(\cdot, \tau)$ for any $\tau \in T$. Then,*

1. *for each $\delta > \max_{i,t} \delta^*_{i,t}$, the planner is intergenerationally Pareto and strongly non-dictatorial;*

2. *for each δ such that for some i, t, $\delta < \delta^*_{i,t}$, there exists some $T^* > 0$ such that if $T \geq T^*$, the planner is not intergenerationally Pareto.*

Proof. **Part I** This part is similar to Part I of Theorem 15. First, we prove a lemma for one-individual aggregation.

Lemma 18. *Assume that $N = \{i\}$. Suppose each generation-t individual i's discounting utility function has instantaneous utility functions $u_i(\cdot, \tau)$ and a discount function $\delta_{i,t}$ such that (A2) and (A3) hold. Let the planner's instantaneous utility function be $u_i(\cdot, \tau)$ for any $\tau \in T$. For any $\delta > \max_t \delta^*_{i,t}$, the planner is intergenerationally Pareto and strongly non-dictatorial.*

Proof. We want to show that for any $\delta > \max_{i,t} \delta^*_{i,t}$, there exists a finite sequence of positive numbers $(\omega_t(i, s))_{t \in T, s \geq t}$ such that

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau - t} u(p_\tau, \tau) = \sum_{s=t}^{T} \omega_t(i, s) U_{i,s}(\mathbf{p}).$$

for each $t \in T$. Given any $\delta > \max_t \delta^*_{i,t}$, we can construct $(\omega_t(i,s))_{t\in T, s\geq t}$ according to the following formula recursively:

$$\omega_t(i,s) = \begin{cases} 1, & \text{if } s = t, \\ \sum_{\tau=t}^{s-1}[\delta \cdot \delta_{i,\tau}(s-1-\tau) - \delta_{i,\tau}(s-\tau)]\omega_t(i,\tau), & \text{if } s > t. \end{cases} \tag{II.29}$$

Note that by assuming $\delta > \max_t \delta^*_{i,t}$, $\omega_t(i,s) > 0$ for any $s \geq t$ and $t \in T$. Then,

$$\begin{aligned} U_t(\mathbf{p}) &= \sum_{s=t}^{T} \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{s=t}^{T} \omega_t(i,s) \sum_{\tau=s}^{T} \delta_{i,s}(\tau - s) u(p_\tau, \tau) \\ &= \sum_{\tau=t}^{T} \left(\sum_{s=t}^{\tau} \delta_{i,s}(\tau - s)\omega_t(i,s) \right) u(p_\tau, \tau). \end{aligned}$$

We want to prove that $U_t(p) = \sum_{\tau=t}^{T} \delta^{\tau-t} u(p_\tau, \tau)$. Clearly for $\tau = t$, $\sum_{s=t}^{\tau} \delta_{i,s}(\tau - s)\omega_t(i,s) = \omega_t(i,t) = 1 = \delta^0$. Suppose for some $\tau \geq t$, we have proven that $\sum_{s=t}^{\tau} \delta_{i,s}(\tau - s)\omega_t(i,s) = \delta^{\tau-t}$. We want to prove that for $\tau + 1$,

$$\sum_{s=t}^{\tau+1} \delta_{i,s}(\tau + 1 - s)\omega_t(i,s) = \delta^{\tau-t+1}. \tag{II.30}$$

To prove (II.30), we only need to notice that according to (II.29),

$$\begin{aligned} \sum_{s=t}^{\tau+1} \delta_{i,s}(\tau + 1 - s)\omega_t(i,s) &= \omega_t(i,\tau+1) + \sum_{s=t}^{\tau} \delta_{i,s}(\tau + 1 - s)\omega_t(i,s) \\ &= \sum_{s=t}^{\tau} [\delta\delta_{i,s}(\tau - s) - \delta_{i,s}(\tau + 1 - s)]\omega_t(i,s) + \sum_{s=t}^{\tau} \delta_{i,s}(\tau + 1 - s)\omega_t(i,s) \\ &= \delta \cdot \sum_{s=t}^{\tau} \delta_{i,s}(\tau - s)\omega_t(i,s) = \delta^{\tau-t+1}. \end{aligned}$$

By induction, we know that $\sum_{s=t}^{\tau} \delta_{i,s}(\tau - s)\omega_t(i,s) = \delta^{\tau-t}$ for all $\tau \geq t$, and hence $U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u_i(p_\tau, \tau)$. □

Next, for any social discount factor $\delta > \max_i \max_t \delta^*_{i,t}$, we can find $(\omega_{i,t}(s))_{t\in T, i\in N, s\geq t}$ such that

$$\sum_{s=t}^{T} \omega_{i,t}(s) U_{i,s}(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau-t} u_i(p_\tau, \tau)$$

for each $i \in N$. Then, we know that

$$U_t(\mathbf{p}) = \sum_{i=1}^{N}\sum_{s=t}^{T} \lambda_i \omega_t(i,s) U_{i,s}(\mathbf{p}) = \sum_{i=1}^{N}\sum_{\tau=t}^{T} \delta^{\tau-t}\lambda_i u_i(p_\tau)$$
$$= \sum_{\tau=t}^{T} \delta^{\tau-t} \sum_{i=1}^{N} \lambda_i u_i(p_\tau,\tau) = \sum_{\tau=t}^{T} \delta^{\tau-t} u(p_\tau,\tau).$$

Part II We prove it by contradiction. Suppose there exists an intergenerationally Pareto planner with social discount factor $\delta < \delta^*_{i,t}$ for some $i = i^*$ and $t = t^*$. By intergenerational Pareto, for each $t \in T$, there exists a finite sequence of nonnegative numbers $(\omega_t(i,s))_{i\in N, s\geq t}$ such that the following equality holds

$$\delta^{\tau-t} u(p_\tau,\tau) = \sum_{i=1}^{N}\sum_{s=t}^{\tau} \omega_t(i,s)\delta_{i,s}(\tau-s)u_i(p_\tau,\tau) \tag{II.31}$$

for any $\tau \geq t$. When $\tau = t$, the above equation reduces to

$$u(p_\tau,\tau) = \sum_{i=1}^{N} \omega_\tau(i,\tau) u_i(p_\tau,\tau) \tag{II.32}$$

for any $\tau \in T$.

Since $u(\cdot,\tau) = \sum_{i\in N} \lambda_i u_i(\cdot,\tau)$ for any $\tau \in T$ and $(u_i(\cdot,\tau))_{i\in N}$ is linearly independent, $\omega_t(i,t) = \lambda_i > 0$, for any i and t. Multiply $\delta^{\tau-t}$ to both sides of equation (II.32) and combine it with equation (II.31). We obtain

$$\sum_{i=1}^{N} \omega_\tau(i,\tau)\delta^{\tau-t} u_i(p_\tau,\tau) = \sum_{i=1}^{N}\sum_{s=t}^{\tau} \omega_t(i,s)\delta_{i,s}(\tau-s)u_i(p_\tau,\tau).$$

Since $(u_i(\cdot,\tau))_{i=1}^N$ is linearly independent, the above equation is equivalent to

$$\omega_\tau(i,\tau)\delta^{\tau-t} u_i(p_\tau,\tau) = \sum_{s=t}^{\tau} \omega_t(i,s)\delta_{i,s}(\tau-s)u_i(p_\tau,\tau)$$

for any $i \in N$, $t \in T$, and $\tau \geq t$.

Let $i = i^*$ and $t = t^*$, and rearrange the above equations. We have

$$\begin{aligned}\delta^{\tau-t^*} &= \frac{\sum_{s=t^*}^{\tau} \omega_{t^*}(i^*, s)\delta_{i^*,s}(\tau - s)}{\omega_\tau(i^*, \tau)} \\ &= \frac{\omega_{t^*}(i^*, t^*)\delta_{i^*,t^*}(\tau - t^*) + \sum_{s=t^*+1}^{\tau} \omega_{t^*}(i^*, s)\delta_{i^*,s}(\tau - s)}{\omega_\tau(i^*, \tau)} \\ &\geq \frac{\lambda_i^* \cdot \delta_{i^*,t^*}(\tau - t^*)}{\lambda_i^*} = \delta_{i^*,t^*}(\tau - t^*)\end{aligned} \tag{II.33}$$

for any $\tau > t^*$. However, we also know that $\delta < \lim_{\tau\to\infty} \sqrt[\tau]{\delta_{i^*,t^*}(\tau)}$, there exists T^* such that for any $\tau \geq T^*$, $\delta < \sqrt[\tau]{\delta_{i^*,t^*}(\tau)}$, which contradicts (II.33). □

When we assume $u(\cdot, \tau) = \sum_{i\in N} \lambda_i u_i(\cdot, \tau)$, we have assumed that λ_i's do not depend on τ. In the social choice literature, some economists have argued that with normalized individual utility functions, equal utilitarian weights should be used (see Karni (1998), Dhillon and Mertens (1999), and Segal (2000)). To some extent, this is consistent with our assumption that λ_i's do not change over time, although in our case, λ_i's may not be $1/N$. In general, one may want λ_i's to depend on time. In that case, the fact that the planner's discount function is exponential will impose restrictions on how λ_i's may change over time.

Strengthening Intergenerational Pareto

The premise of intergenerational Pareto requires that the current generation and future generations reach a consensus. A natural way to strengthen intergenerational Pareto may be to require that the planner prefer one consumption sequence over another if more than a certain fraction of current- and future-generation individuals agree.[17] However, in this case, how the planner aggregates individual preferences may differ somewhat from utilitarian aggregation.

Therefore, we strengthen intergenerational Pareto in the following simple way without deviating from standard utilitarianism. Let $I \subset N \times T$ be an arbitrary subset of individuals across generations. Let us weaken the premise of intergenerational Pareto by requiring that the planner prefer a consumption sequence $\mathbf{p}$ to $\mathbf{q}$ whenever individuals in I agree. Intergenerational Pareto and the strongly non-dictatorial property are adapted as follows.

Definition 19. The planner's preference $(\succsim_t)_{t\in T}$ is I-intergenerationally Pareto if for any consumption sequences $\mathbf{p}, \mathbf{q} \in \Delta(X)^T$, in each period $t \in T$, $\mathbf{p} \succsim_{i,s} \mathbf{q}$ for all $(i, s) \in I$ with $s \geq t$ implies $\mathbf{p} \succsim_t \mathbf{q}$, and $\mathbf{p} \succ_{i,s} \mathbf{q}$ for all $(i, s) \in I$ with $s \geq t$ implies $\mathbf{p} \succ_t \mathbf{q}$.

[17] This strengthening can certainly be applied to current-generation Pareto as well.

Definition 20. We say that the planner is I-strongly non-dictatorial if for each $t \in T$,

$$U_t(\mathbf{p}) = f_t\left(U_{1,t}(\mathbf{p}), \ldots, U_{1,T}(\mathbf{p}), U_{2,t}(\mathbf{p}), \ldots, U_{2,T}(\mathbf{p}), \ldots, U_{N,T}(\mathbf{p})\right)$$

for some function f_t that is (strictly) increasing in $U_{i,s}$ for any $(i,s) \in I$.

It is straightforward to show that under I-intergenerational Pareto, the planner's utility function can be written as a weighted sum of the utility functions of individuals in I. Below, we show that under some assumption about I, positive results can still be established after strengthening intergenerational Pareto.

The following example shows why we need an additional assumption. Suppose $N = 2$ and individual instantaneous utility functions, u_1 and u_2, are linearly independent. Assume that $I = \{(2,1),(1,2)\}$; that is, the planner will give generation-1 individual 1 and generation-2 individual 2 zero weights. Then, the somewhat trivial negative result, as in Example 16, appears again. To see this, note that in period 1, the planner's instantaneous utility function for period-1 consumption must be equal to u_2, because only generation-1 individuals care about period-1 consumption and generation-1 individual 1 has been ignored. We have assumed that the planner has an EDU function, in which her instantaneous utility function never changes. Now, first, in period 1, the planner's instantaneous utility function for period-2 consumption is a strict convex combination of u_1 and u_2, which must differ from u_1; second, in period 2, following the same logic, the planner's instantaneous utility function for period-2 consumption must be equal to u_1, which is again different from u_1. Therefore, it is hopeless to derive any positive result.

The theorem below imposes a simple assumption to avoid the example above, which turns out to be strong enough for us to establish a positive result. For each $t \in T$, let $I_t := \{i \in N : (i,t) \in I\}$ be the set of generation-t individuals who may not be ignored by the planner, and let $\mathcal{I} := \bigcup_{t \in T} I_t$.

Theorem 21. *Suppose $I \subset N \times T$, and each generation-t individual i's discounting utility function has an instantaneous utility function $u_i \in \{u^\theta\}_{\theta=1}^{\Theta}$ for some linearly independent Θ-tuple of instantaneous utility functions $(u^\theta)_{\theta=1}^{\Theta}$, and has a discount function δ_i such that (A2) and (A3) hold. Assume that $co(\{u_i\}_{i \in I_t}) = co(\{u^\theta\}_{\theta=1}^{\Theta})$ for any $t \in T$. Let the planner's instantaneous utility function u be a strict convex combination of $(u_i)_{i \in I_t}$. Then,*

1. *for each $\delta > \max_{\mathcal{I}} \delta_i^*$, the planner is I-intergenerationally Pareto and I-strongly non-dictatorial;*

2. *for each $\delta < \min_{\mathcal{I}} \delta_i^*$, there exists some $T^* > 0$ such that if $T \geq T^*$, the planner is not I-intergenerationally Pareto.*

Proof. **Part I** With an abuse of notation, let $\Theta := \{1, \ldots, \Theta\}$. For each $\theta \in \Theta$, let $I^\theta := \{i \in N : u_i = u^\theta\}$, which is the set of i's whose instantaneous utility function is u^θ. For each $\theta \in \Theta$ and $t \in T$, let $I_t^\theta := \{i \in I^\theta : (i,t) \in I\}$ be the set of generation-t individuals who may not be ignored by the planner and whose instantaneous utility function is u^θ. Let $\mathcal{I}^\theta := \bigcup_{t \in T} I_t^\theta$.

We prove this part in four steps. First, we aggregate individuals in each I_t^θ into a new "family" θ. Each generation-t family θ has instantaneous utility function $u^\theta(\cdot)$ and the following discount function:

$$\delta_t^\theta(\tau) = \frac{1}{|I_t^\theta|} \sum_{i \in I_t^\theta} \delta_i(\tau);$$

that is, if a generation-t individual i may not be ignored by the planner, his discount function $\delta_i(\cdot)$ enters family θ's generation-t discount function $\delta_t^\theta(\cdot)$ with a weight equal to that of other generation-t individual(s) in I_t^θ. Note that generation-t families' discount functions may change as t changes.

Next, we prove a lemma on one-family aggregation that is similar to Lemma 18.

Lemma 22. *Assume $\Theta = \{\theta\}$. Suppose each generation-t family θ's discounting utility function has an instantaneous utility function $u^\theta(\cdot)$ and a discount function $\delta_t^\theta(\cdot)$. Let the planner's instantaneous utility function be $u^\theta(\cdot)$. For any $\delta > \max_{i \in \mathcal{I}^\theta} \delta_i^*$, the planner is intergenerationally Pareto and strongly non-dictatorial.*

Proof. We want to show that for any $\delta > \max_{i \in \mathcal{I}^\theta} \delta_i^*$, there exists a finite sequence of positive numbers $(\omega_t^\theta(s))_{t \in T, s \geq t}$ such that

$$U_t(\mathbf{p}) = \sum_{\tau=t}^{T} \delta^{\tau - t} u^\theta(p_\tau) = \sum_{s=t}^{T} \omega_t^\theta(s) U_s^\theta(\mathbf{p})$$

for each $t \in T$. Given any $\delta > \max_{i \in \mathcal{I}^\theta} \delta_i^*$, we can construct $(\omega_t^\theta(s))_{t \in T, s \geq t}$ according to the following formula recursively:

$$\omega_t^\theta(s) = \begin{cases} 1, & \text{if } s = t, \\ \sum_{\tau=t}^{s-1} [\delta \cdot \delta_\tau^\theta(s - 1 - \tau) - \delta_\tau^\theta(s - \tau)] \omega_t^\theta(\tau), & \text{if } s > t. \end{cases} \tag{II.34}$$

Note that if $\delta > \max_t \max_\tau \frac{\delta_t^\theta(\tau+1)}{\delta_t^\theta(\tau)}$, then $\omega_t^\theta(s) > 0$ for any $s \geq t$ and $t \in T$. We also

know that

$$\begin{aligned}\frac{\delta_t^\theta(\tau+1)}{\delta_t^\theta(\tau)} &= \frac{\sum_{i\in I_t^\theta}\delta_i(\tau+1)}{\sum_{i\in I_t^\theta}\delta_i(\tau)} = \frac{\sum_{i\in I_t^\theta}\delta_i(\tau)\frac{\delta_i(\tau+1)}{\delta_i(\tau)}}{\sum_{i\in I_t^\theta}\delta_i(\tau)} \leq \frac{\sum_{i\in I_t^\theta}\delta_i(\tau)\max_{i\in I_t^\theta}\frac{\delta_i(\tau+1)}{\delta_i(\tau)}}{\sum_{i\in I_t^\theta}\delta_i(\tau)} \\ &\leq \max_{i\in I_t^\theta}\frac{\delta_i(\tau+1)}{\delta_i(\tau)} \leq \max_{i\in I_t^\theta}\delta_i^* \leq \max_{i\in\mathcal{I}^\theta}\delta_i^*.\end{aligned}$$

Therefore, $\max_t \max_\tau \frac{\delta_t^\theta(\tau+1)}{\delta_t^\theta(\tau)} \leq \max_{i\in\mathcal{I}^\theta}\delta_i^*$. Hence, by assuming $\delta > \max_{i\in\mathcal{I}^\theta}\delta_i^*$, $\omega_t^\theta(s) > 0$ for any $s \geq t$ and $t \in T$. The rest of the proof is the same as in Lemma 18. □

Thus, for any social discount factor $\delta > \max_{\theta\in\Theta}\max_{i\in\mathcal{I}^\theta}\delta_i^*$, we can find $(\omega_t^\theta(s))_{t\in T,\theta\in\Theta,s\geq t}$ such that

$$\sum_{s=t}^{T}\omega_t^\theta(s)U_s^\theta(\mathbf{p}) = \sum_{\tau=t}^{T}\delta^{\tau-t}u^\theta(p_\tau)$$

for each $\theta \in \Theta$. Consider any positive numbers $(\lambda^\theta)_{\theta=1}^\Theta$ such that $\sum_{\theta=1}^\Theta \lambda^\theta = 1$. Together with the weights $(\omega_t^\theta(s))_{t\in T,\theta\in\Theta,s\geq t}$ we have found above, the planner's utility function becomes

$$\begin{aligned}U_t(\mathbf{p}) &= \sum_{\theta\in\Theta}\sum_{s=t}^{T}\lambda^\theta\omega_t^\theta(s)U_s^\theta(\mathbf{p}) = \sum_{\theta\in\Theta}\sum_{\tau=t}^{T}\lambda^\theta\delta^{\tau-t}u^\theta(p_\tau) \\ &= \sum_{\tau=t}^{T}\delta^{\tau-t}\sum_{\theta\in\Theta}\lambda^\theta u^\theta(p_\tau) = \sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau),\end{aligned} \tag{II.35}$$

in which $u(p_\tau) = \sum_{\theta\in\Theta}\lambda^\theta u^\theta(p_\tau)$ can be any strict convex combination of $(u^\theta)_{\theta\in\Theta}$.

Lastly, we back out the weights $(\omega_t(i,s))_{t\in T,i\in N,s\geq t}$ and show that the planner has an EDU function, is I-intergenerationally Pareto, and is I-strongly non-dictatorial under these weights. We construct $(\omega_t(i,s))_{t\in T,i\in N,s\geq t}$ according to the following formula:

$$\omega_t(i,s) = \begin{cases} 0, & \text{if } (i,s) \notin I, \\ \lambda^\theta\frac{1}{|I_s^\theta|}\omega_t^\theta(s) > 0, & \text{if } (i,s) \in I. \end{cases}$$

Then,

$$\begin{aligned}
\sum_{s=t}^{T}\sum_{i=1}^{N}\omega_t(i,s)U_{i,s}(\mathbf{p}) &= \sum_{s=t}^{T}\sum_{i=1}^{N}\omega_t(i,s)\sum_{\tau=s}^{T}\delta_i(\tau-s)u_i(p_\tau)\\
&= \sum_{s=t}^{T}\sum_{\theta\in\Theta}\sum_{i\in I_s^\theta}\lambda^\theta\frac{1}{|I_s^\theta|}\omega_t^\theta(s)\sum_{\tau=s}^{T}\delta_i(\tau-s)u_i(p_\tau)\\
&= \sum_{s=t}^{T}\sum_{\theta\in\Theta}\lambda^\theta\omega_t^\theta(s)\sum_{\tau=s}^{T}\sum_{i\in I_s^\theta}\frac{1}{|I_s^\theta|}\delta_i(\tau-s)u_i(p_\tau)\\
&= \sum_{s=t}^{T}\sum_{\theta\in\Theta}\lambda^\theta\omega_t^\theta(s)\sum_{\tau=s}^{T}\delta_s^\theta(\tau-s)u^\theta(p_\tau)\\
&= \sum_{s=t}^{T}\sum_{\theta\in\Theta}\lambda^\theta\omega_t^\theta(s)U_s^\theta(\mathbf{p}) = U_t(\mathbf{p}) = \sum_{\tau=t}^{T}\delta^{\tau-t}u(p_\tau).
\end{aligned}$$

The first equality follows from the definition of $U_{i,s}$. The second equality follows the construction of $(\omega_t(i,s))_{t\in T,i\in N,s\geq t}$. The fourth equality follows the construction of $\delta_s^\theta(\cdot)$. The fifth equality follows from the definition of U_s^θ. The last two equalities follow equation (II.35).

Part II We prove it by contradiction. Suppose there exists an I-intergenerationally Pareto planner with social discount factor $\delta < \min_{i\in\mathcal{I}}\delta_i^*$. By I-intergenerationally Pareto, there exists a finite sequence of nonnegative weights $(\omega_t(i,s))_{t\in T,i\in N,s\geq t}$ such that the following equality holds:

$$\delta^{\tau-t}u(p_\tau) = \sum_{s=t}^{\tau}\sum_{i\in I_s}\omega_t(i,s)\delta_i(\tau-s)u_i(p_\tau) \tag{II.36}$$

for each $t\in T$ and $\tau\geq t$. Combining equation (II.36) with the normalization assumption,

$$\delta^{\tau-t} = \sum_{s=t}^{\tau}\sum_{i\in I_s}\omega_t(i,s)\delta_i(\tau-s) \geq \sum_{i\in I_t}\omega_t(i,s)\delta_i(\tau-s) \tag{II.37}$$

for each $t\in T$ and $\tau\geq t$.

We assume that $\arg\min_{i\in\mathcal{I}}\delta^* = \{i^*\}$. The proof can be easily extended to the case with multiple minima. The following two claims must hold:

1. $i^*\in\mathcal{I}$; that is, there exists $t^*\in T$ such that $i^*\in I_{t^*}$.

2. There exists T_1 such that for any $\tau\geq\max\{T_1,t^*\}$, $\delta_{i^*}(\tau-t^*)\leq\delta_i(\tau-t^*)$ for any $i\in\mathcal{I}$.

Consider the period-t^* planner. Let $t=t^*$ in equation (II.37), and suppose $\tau\geq$

$\max\{T_1, t^*\}$. We have

$$\begin{aligned}\delta^{\tau-t^*} &= \sum_{s=t^*}^{\tau}\sum_{i\in I_s}\omega_{t^*}(i,s)\delta_i(\tau-s) \geq \sum_{i\in I_t^*}\omega_{t^*}(i,t^*)\delta_i(\tau-t^*) \\ &\geq \sum_{i\in I_t^*}\omega_{t^*}(i,t^*)\delta_{i^*}(\tau-t^*) \geq \delta_{i^*}(\tau-t^*)\end{aligned}$$

However, we know that $\delta < \delta_{i^*}^*$. Then, there exists T_2 such that for any $\tau \geq T_2$, $\delta < \sqrt[\tau]{\delta_{i^*}^*}$. Therefore, if $T \geq \max\{T_1, T_2, t^*\}$, there must be a contradiction. □

Note that we assume $\text{co}(\{u_i\}_{i\in I_t})$ =$\text{co}(\{u^\theta\}_{\theta=1}^{\Theta})$ for any $t \in T$. This is because we want $\text{co}(\{u_i\}_{i\in I_t})$ to remain constant across t to rule out the example we discuss before the theorem, and we want to assume that there is no redundant type.

The theorem seems different from our previous results that have only one cutoff for the social discount factor, but in fact it has a one-cutoff version that is similar to our previous positive results. However, the expression of the cutoff will become rather complicated.[18] The current version is easier to understand, and clearly shows that if the social discount factor is higher than the highest long-run discount factor among individuals who are not ignored in some generation, then we know that the planner is intergenerationally Pareto and strongly non-dictatorial. Again, this is not the only way to establish positive results. If the planner's instantaneous utility function is allowed to vary in a general way by taking the form of $u_t(\cdot, \tau)$, then the additional assumption we need can be weaker.

Utilitarianism and Long-Run Social Discounting

The main question of this paper is, if a planner has an EDU function, under what conditions is she intergenerationally Pareto/utilitarian and strongly non-dictatorial? The fact that an intergenerationally Pareto/utilitarian planner has an EDU function certainly imposes restrictions on how the planner may aggregate individual preferences. On the one hand, economists often assume that a planner has an EDU function, and there are many reasons to believe that this is normatively appealing. Therefore, understanding the answer to our main question is important.

On the other hand, there are other ways to examine the planner's aggregation problem. For example, sometimes economists may believe that the planner's utility function should be equal to the simple average of individuals' discounting utility functions. However, because

[18]The cutoff for the social discount factor in the one-cutoff version should take the maximum across types and periods, and then for each type in each period, take the minimal individual long-run discount factor across all individuals who have the desired type and are not ignored in that period.

it is unlikely that the planner's discount function is exponential in this case, a choice about what to assume for the planner must be made.

A natural question arises: If we now want to allow the planner to aggregate individual preferences in a flexible way—in other words, we only require that the planner be intergenerationally Pareto/utilitarian and strongly non-dictatorial and do not require that her utility function be an EDU function—what insight from our main findings remains true? The following result shows that under this different requirement, the planner's "discount factor" should still be higher than the most patient individual's long-run discount factor. The result assumes that $T = +\infty$. Some notations and defintions for the case with $T = +\infty$ can be found in Section II.7.8.

Theorem 23. *Suppose $T = +\infty$, each generation-t individual i's discounting utility function has an instantaneous utility function u_i and a discount function δ_i such that (A2) and (A3) hold, and the planner's utility function in any period $t \in T$ is $U_t = \sum_{\tau=t}^{\infty} \delta_t(\tau - t) u_t(p_\tau, \tau)$ for some discount function δ_t and (normalized) instantaneous utility function $(u_t(\cdot, \tau))_{\tau \geq t}$ such that $\delta_t^* = \lim_{\tau \to \infty} \frac{\delta_t(\tau+1)}{\delta_t(\tau)} = \lim_{\tau \to \infty} \sqrt[\tau]{\delta_t(\tau)}$ exists. If the planner is intergenerationally utilitarian and strongly non-dictatorial, $\delta_t^* \geq \max_i \delta_i^*$.*

Proof. Since $U_t = \sum_{i=1}^{N} \sum_{s=t}^{\infty} \omega_t(i, s) U_{i,s}$, we know that

$$\delta_t(\tau - t) u_t(p_\tau, \tau) = \sum_{i=1}^{N} \sum_{s=t}^{\tau} \omega_t(i, s) \delta_i(\tau - s) u_i(p_\tau) \tag{II.38}$$

for any $t \in T$ and $\tau \geq t$. Let $p_\tau = x^*$ in equation (II.38). We have

$$\delta_t(\tau - t) = \sum_{i=1}^{N} \sum_{s=t}^{\tau} \omega_t(i, s) \delta_i(\tau - s) \geq \sum \omega_t(i, t) \delta_i(\tau - t) \geq \omega_t(i^*, t) \delta_{i^*}(\tau - t), \tag{II.39}$$

in which $i^* := \arg\max_i \delta_i^*$. Let τ in (II.39) go to infinity. We have $\delta_t^* \geq \max_i \delta_i^*$.

Thus, if the planner is intergenerationally Pareto and strongly non-dictatorial, her long-run discount factor should again be higher than the most patient individual's long-run discount factor. □

II.8.2 An Alternative Interpretation of Intergenerational Pareto and a Result with Quasi-Hyperbolic Discounting

We say that the generation-t individual i has a *quasi-hyperbolic discounting utility* (QHDU) function if his discount function satisfies

$$\delta_i(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ \beta_i \delta_i^\tau, & \text{if } \tau \in \{1, \dots, T-1\}, \end{cases}$$

for some $\beta_i \in (0, 1]$ and $\delta_i > 0$. In the literature of time inconsistency, economists sometimes ignore the β_i parameter and use an EDU function with a discount factor δ_i as the welfare criterion of individual i who has a QHDU function. The intuition is that because β_i is the cause of time inconsistency, β_i should not enter the welfare criterion. We show how our analysis provides some foundation for this practice.[19]

Consider Corollary 8. If we interpret the generation-$(t+1)$ individual i in our model as the future self of the generation-t individual i, Corollary 8 provides some foundation for the use of this welfare criterion. Assume that individual i is the only individual ($N = 1$) and has a quasi-hyperbolic discount function. According to Corollary 8, we immediately know that any EDU function with a discount factor that is (strictly) greater than δ_i is a welfare criterion that is consistent with intergenerational Pareto; that is, if individual i in every period t agrees that one consumption sequence is better than another, the welfare criterion says that the utility of the former is greater than the latter.

The following result is stronger than Corollary 8. It shows that δ_i is indeed the smallest discount factor such that the corresponding EDU function is consistent with intergenerational Pareto.

Proposition II.5. *Suppose each generation-t individual i has a QHDU function with an instantaneous utility function u, $\beta_i \in (0, 1)$, and $\delta_i \in (0, 1)$. Then,*

1. *for each $\delta \geq \min_i \delta_i$, the planner is intergenerationally Pareto and strongly non-dictatorial;*

2. *for each $\delta < \min_i \delta_i$, there exists some $T^* > 0$ such that if $T \geq T^*$, the planner is not intergenerationally Pareto.*

Proof. The second part follows from Theorem 6. We only prove the first part.

[19] Recent papers by Drugeon and Wigniolle (2017) and Galperti and Strulovici (2017) introduce results similar to the one we present below.

Lemma 24. *Assume that $N = \{i\}$. Suppose individual i has a QHDU function with parameters $\beta_i \in (0,1)$, $\delta_i \in (0,1)$, and u. Then, there exists a cutoff $\delta(T)$ for each T such that the planner is intergenerationally Pareto and strongly non-dictatorial if and only if $\delta > \delta(T)$. In addition, $\delta(T)$ is (strictly) increasing with a limit δ_i.*

Proof. The planner is intergenerationally Pareto and strongly non-dictatorial if and only if there exists a finite sequence of positive weights $(\omega_t(i,s))_{t\in T, s\geq t}$ such that the following equation holds:

$$\omega_t(i,\tau)u(p_\tau) + \sum_{s=t}^{\tau-1} \omega_{i,t}(s)\beta_i\delta_i^{\tau-s}u(p_\tau) = \delta^{\tau-t}u(p_\tau) \tag{II.40}$$

for any $t \in T$ and $\tau \geq t$. We can solve $(\omega_t(i,s))_{t\in T, s\geq t}$ from (II.40) as follows:

$$\omega_t(i,t+m) = \begin{cases} 1, & \text{if } m = 0, \\ \delta^m - \dfrac{\beta_i}{1-\beta_i}\displaystyle\sum_{h=1}^{m}(1-\beta_i)^h\delta_i^h\delta^{m-h}. & \text{if } 1 \leq m \leq T-t. \end{cases} \tag{II.41}$$

Note that $\omega_t(i,t) = 1 > 0$, and the planner is intergenerationally Pareto and strongly non-dictatorial if and only if $(\omega_t(i,t+m))_{t\in T, 1\leq m\leq T-t}$ is positive.

We can rewrite the second equation of (II.41) as $\omega_t(i,t+m) = F_m(\delta|\beta_i,\delta_i)$, in which F is a degree-m polynomial of a single indeterminate δ with parameters β_i, δ_i. Define

$$S(\beta_i,\delta_i,T) := \{\delta \in \mathbb{R}_+ : F_m(\delta|\beta_i,\delta_i) > 0 \text{ for any } 1 \leq m \leq T-1\}.$$

Therefore, the planner's preference is intergenerationally Pareto and strongly non-dictatorial if and only if $\delta \in S(\beta_i,\delta_i,T)$.

We want to show that $S(\beta_i,\delta_i,T)$ is an interval that (strictly) shrinks to $[\delta_i,+\infty)$ as T increases. First, we prove that there exists a unique root/cutoff $x_m \in (0,\delta_i]$ for $F_m(\delta|\beta_i,\delta_i)$ such that $F_m(x_m|\beta_i,\delta_i) = 0$, $F_m(\delta|\beta_i,\delta_i) < 0$ for $\delta < x_m$, and $F_m(\delta|\beta_i,\delta_i) > 0$ for $\delta > x_m$. We know that $F_m(0|\beta_i,\delta_i) = -(1-\beta_i)^{m-1}\delta_i^m < 0$, $F_m(\delta_i|\beta_i,\delta_i) = (1-\beta_i)^m\delta_i^m > 0$, and F_m is continuous. Therefore, the existence of x_m is guaranteed by Bolzano's theorem.

Also note that the function $G_m(\delta|\beta_i,\delta_i) := \delta^{-m}F_m(\delta|\beta_i,\delta_i)$ has the same root as $F_m(\delta|\beta_i,\delta_i)$, and $G_m(\delta|\beta_i,\delta_i)$ is (strictly) increasing in δ because

$$\frac{dG_m(\delta)}{d\delta} = \frac{\beta_i}{1-\beta_i}\sum_{k=1}^{m} k\frac{(1-\beta_i\delta_i)^k}{\delta^{k+1}} > 0.$$

By Rolle's theorem, there cannot be more than one root. Hence, the uniqueness is proved.

Second, we prove that the cutoff sequence $(x_m)_m$ is (strictly) increasing and converges

to δ_i. Noting that

$$G_{m+1}(\delta|\beta_i,\delta_i) - G_m(\delta|\beta_i,\delta_i) = -\frac{\beta_i}{1-\beta_i}\left[\frac{(1-\beta_i)\delta_i}{\delta}\right]^{m+1} < 0,$$

we have $G_{m+1}(x_m|\beta_i,\delta_i) - G_m(x_m|\beta_i,\delta_i) < 0$. By the definition of $(x_m)_m$, $G_m(x_m|\beta_i,\delta_i) = G_{m+1}(x_{m+1}|\beta_i,\delta_i) = 0$. Therefore, $G_{m+1}(x_m|\beta_i,\delta_i) < G_m(x_m|\beta_i,\delta_i) = G_{m+1}(x_{m+1}|\beta_i,\delta_i)$. We also know that $G_m(\delta|\beta_i,\delta_i)$ is (strictly) increasing. Hence, $x_{m+1} > x_m$. Now that $(x_m)_m$ is bounded and (strictly) increasing, the convergence follows from the monotone convergence theorem.

The only remaining part is to prove that the limit of the cutoff sequence is δ_i. Suppose $\lim_{m\to\infty} x_m = x$. Then, $x_m < x$ for all $m > 1$. Since $G_m(\delta|\beta_i,\delta_i)$ is (strictly) increasing, we have

$$\begin{aligned}
&G_m(x_m|\beta_i,\delta_i) < G_m(x|\beta_i,\delta_i)\\
\Leftrightarrow\, &0 < 1 - \frac{\beta_i}{1-\beta_i}\sum_{h=1}^{m}(1-\beta_i)^h\delta_i^h x^{-h}\\
\Leftrightarrow\, &\frac{\beta_i}{1-\beta_i}\sum_{h=1}^{m}(1-\beta_i)^h\delta_i^h x^{-h} < 1\\
\Leftrightarrow\, &\sum_{h=1}^{m}\left[\frac{(1-\beta_i)\delta_i}{x}\right]^h < \frac{1-\beta_i}{\beta_i}
\end{aligned} \tag{II.42}$$

for any $m > 1$.

Given that $\frac{(1-\beta_i)\delta_i}{x} > 0$, we must have $\frac{(1-\beta_i)\delta_i}{x} < 1$; otherwise, $\sum_{h=1}^{m}\left[\frac{(1-\beta_i)\delta_i}{x}\right]^h$ diverges as m increases. Now, let m in (II.42) go to infinity. We have

$$\begin{aligned}
&\sum_{h=1}^{+\infty}\left[\frac{(1-\beta_i)\delta_i}{x}\right]^h \le \frac{1-\beta_i}{\beta_i}\\
\Leftrightarrow\, &\frac{(1-\beta_i)\delta_i}{x}\frac{1}{1-\frac{(1-\beta_i)\delta_i}{x}} \le \frac{1-\beta_i}{\beta_i}\\
\Leftrightarrow\, &\delta_i \le x.
\end{aligned} \tag{II.43}$$

In addition, since $x_m < \delta_i$ for all m, we have $x \le \delta_i$. Therefore, $x = \delta_i$. □

Lemma 24 states that for any finite T, in each period t, the planner can aggregate each individual i from the t^{th} generation to the T^{th} generation so that the aggregated utility function is an EDU function with a discount factor that is slightly below δ_i. Then, we can apply the if part of Proposition II.2 for N exponential discounting individuals, and obtain a

social discount factor $\delta \geq \min_i \delta_i$. □

When $T = +\infty$, we can assume in Proposition II.4 that individuals have QHDU functions and obtain a similar result.

II.8.3 The Case with Backward Discounting

The result we introduce below shows that if individuals exponentially forward and backward discount consumption, our main results continue to hold. Before proceeding, it should be noted that backward discounting has no revealed-preference foundation. Whenever we observe an individual choosing, the past is sunk; there are no choices (yet) that allow the individual to alter the past. Therefore, we do not know how individuals think about the past from actual choice data.

However, economists have considered the possibility that individuals backward discount (see Strotz (1955), Caplin and Leahy (2004), and Ray, Vellodi and Wang (2017)). Below, we analyze our aggregation problem with exponential discounting individuals who backward discount. Instead of assuming that $U_{i,t}(\mathbf{p})$ does not depend on past consumption, we assume that the generation-t individual i discounts both past and future by the same discounting factor δ_i.

Definition 25. The generation-t individual i has an exponential forward and backward discounting utility function if his utility function has the following form:

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=1}^{T} \delta_i^{|\tau-t|} u_i(p_\tau), \tag{II.44}$$

in which the discount factor $\delta_i \in (0,1)$, and u_i is the individual i's instantaneous utility function.

Note that the negative result, obviously, would continue to hold if we had assumed that each generation-t individual i's utility function was

$$U_{i,t}(\mathbf{p}) = \sum_{\tau=1}^{T} \delta_i^{\tau-t} u_i(p_\tau).$$

In that case, the individual i's offspring has exactly the same preference as the indivdual i. This is problematic, however, because the generation-2 individual i will value period-1 consumption even more than his own period-2 consumption.

The result below demonstrates that the assumption that the planner has an EDU function and intergenerational Pareto are compatible when individuals exponentially forward

and backward discount consumption. The typical negative result in the literature only considers the planner's aggregation problem in period 1. The following result also focuses on the period-1 aggregation problem to highlight the difference.

Proposition II.6. *Suppose each generation-t individual i has an exponential forward and backward discounting utility function with discount factor δ_i and instantaneous utility function u_i such that $\bar{\delta} := \max_i \delta_i < 1$. Let the planner's instantaneous utility function u be an arbitrary strict convex combination of $(u_i)_{i\in N}$. Then, for each $\delta \in \left(\bar{\delta}, \bar{\delta}^{-1}\right)$, the planner in period 1 is intergenerationally Pareto and strongly non-dictatorial.*

Proof. To prove the proposition, we consider the one-individual case first.

Lemma 26. *Assume that $N = \{i\}$. Suppose each generation-t individual i has an exponential forward and backward discounting utility function with discount factor $\delta_i \in (0, 1)$ and instantaneous utility function u. Then, for each $\delta \in (\delta_i, {\delta_i}^{-1})$, the planner in period 1 is intergenerationally Pareto and strongly non-dictatorial.*

Proof. We want to show that for any $\delta \in (\delta_i, {\delta_i}^{-1})$, there exists a finite sequence of positive weights $\vec{\omega} = (\omega(i,1), \omega(i,2), \cdots, \omega(i,T))$ such that the following equation holds:

$$U_1(\mathbf{p}) = \sum_{\tau=1}^{T} \delta^{\tau-1} u(p_\tau) = \sum_{s=1}^{T} \omega(i,s) U_{i,s}(\mathbf{p}). \tag{II.45}$$

Plugging in $U_1(\mathbf{p})$ and $U_{i,s}(\mathbf{p})$, equation (II.45) becomes

$$\sum_{\tau=1}^{T} \delta^{\tau-1} u(p_\tau) = \sum_{s=1}^{T} \omega(i,s) \sum_{\tau=1}^{T} \delta_i^{|\tau-s|} u(p_\tau) = \sum_{\tau=1}^{T} \sum_{s=1}^{T} \omega(i,s) \delta_i^{|s-\tau|} u(p_\tau); \tag{II.46}$$

that is, for each $\tau \geq 1$,

$$\delta^{\tau-1} = \sum_{s=1}^{T} \omega(i,s) \delta_i^{|s-\tau|}. \tag{II.47}$$

Next, we can rewrite equation (II.47) as follows:

$$A \cdot \vec{\omega} = \vec{\delta}, \tag{II.48}$$

in which $\vec{\delta} = (1, \delta, \delta^2, \cdots, \delta^{T-1})$ and

$$A = \begin{pmatrix} 1 & \delta_i & \delta_i^2 & \dots & \delta_i^{T-1} \\ \delta_i & 1 & \delta_i & \dots & \delta_i^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \delta_i^{T-1} & \delta_i^{T-2} & \delta_i^{T-3} & \dots & 1 \end{pmatrix}.$$

Note that A is invertible. In particular,

$$A^{-1} = \frac{1}{1-{\delta_i}^2} \begin{pmatrix} 1 & -\delta_i & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ -\delta_i & 1+\delta_i^2 & -\delta_i & 0 & & & & \vdots \\ 0 & -\delta_i & 1+\delta_i^2 & -\delta_i & \ddots & & & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\ \vdots & & & \ddots & -\delta_i & 1+\delta_i^2 & -\delta_i & 0 \\ \vdots & & & & 0 & -\delta_i & 1+\delta_i^2 & -\delta_i \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & -\delta_i & 1 \end{pmatrix}.$$

We have $\vec{\omega} = A^{-1} \cdot \vec{\delta}$. If we can show that $\vec{\omega} \gg 0$, the lemma is proved. Showing that $\vec{\omega} \gg 0$ is equivalent to showing that $\omega(i,1) = 1 - \delta_i\delta > 0$, $\omega(i,s) = \delta^{s-2}[-\delta_i + (1+\delta_i^2)\delta - \delta_i\delta^2] > 0$ for $2 \leq s \leq T-1$, and $\omega(i,T) = -\delta_i\delta^{T-2} + \delta^{T-1} > 0$, which can be verified because $\delta \in (\delta_i, {\delta_i}^{-1})$. □

Lemma 26 shows that we can aggregate each individual i from the t^{th} generation to the T^{th} generation into an EDU function with any discount factor δ within $(\delta_i, {\delta_i}^{-1})$. Now we can prove Proposition II.6. For any social discount factor $\delta \in (\bar{\delta}, \bar{\delta}^{-1})$, we can find $(\omega(i,s))_{i\in N, s\geq 1}$ such that

$$\sum_{s=1}^{T} \omega(i,s) U_{i,s}(\mathbf{p}) = \sum_{\tau=1}^{T} \delta^{\tau-1} u_i(p_\tau)$$

for each $i \in N$. Consider any positive numbers $(\lambda_i)_{i\in N}$ such that $\sum_{i\in N} \lambda_i = 1$. Together with the weights $(\omega(i,s))_{i\in N, s\geq 1}$ we have found above, the planner's utility function becomes

$$\begin{aligned} U_1(\mathbf{p}) &= \sum_{i=1}^{N}\sum_{s=1}^{T} \lambda_i \omega_1(i,s) U_{i,s}(\mathbf{p}) = \sum_{i=1}^{N}\sum_{\tau=1}^{T} \delta^{\tau-1}\lambda_i u_i(p_\tau) \\ &= \sum_{\tau=t}^{T} \delta^{\tau-1} \sum_{i=1}^{N} \lambda_i u_i(p_\tau) = \sum_{\tau=t}^{T} \delta^{\tau-1} u(p_\tau), \end{aligned}$$

in which $u(p_\tau) = \sum_{i \in N} \lambda_i u_i(p_\tau)$ is an arbitrary strict convex combination of $(u_i)_{i \in N}$. $\square$

II.8.4 Risk Resolution

The main model's choice domain is $\Delta(X)^T$; that is, in each period, there is a lottery/probability measure over X. In many dynamic economic models with uncertainty, uncertainty resolves over time. Below, we discuss what may change if we let uncertainty resolve over time, maintaining our assumptions about individuals' and the planner's utility functions.

For simplicity, assume that $T = 2$ and $N = 1$. In period 2, the choice object is again a *lottery* over X. Sometimes, it will be called a period-2 lottery. To distinguish between choice objects in the main paper and in this section, here we call X *outcomes* and period-1 choice objects *dynamic lotteries*. A dynamic lottery is a lottery over $X \times \Delta(X)$. For example, with probability $1/2$, a dynamic lottery $\tilde{p}_1$ yields a period-1 outcome $x \in X$ and a period-2 lottery $q_2 \in \Delta(X)$; with probability $1/2$, $\tilde{p}_1$ yields a period-1 outcome x' and a period-2 lottery $r_2 \in \Delta(X)$.

Now, the set of dynamic lotteries is $\Delta(X \times \Delta(X))$, rather than $\Delta(X)^2$.[20] However, $\Delta(X)^2$ can be viewed as a subset of $\Delta(X \times \Delta(X))$ that consists of all dynamic lotteries whose period-2 lotteries are independent of (the realization of) period-1 outcomes.

The following simple example shows in what sense, in period 1, the planner's aggregation problem under $\Delta(X \times \Delta(X))$ is the same as under $\Delta(X)^2$. Continue our example of $\tilde{p}_1, q_2, r_2$ above. Let q_2 be a lottery that yields $y, y' \in X$ with equal probability. Let r_2 be a degenerate lottery that yields $z \in X$. First, consider the generation-1 individual. A natural way to extend our period-1 individual utility function on $\Delta(X)^2$ to the new domain $\Delta(X \times \Delta(X))$ is as follows:

$$\begin{aligned} V_1(\tilde{p}_1) &= \frac{1}{2}(v(x,1) + \delta v(q_2,2)) + \frac{1}{2}(v(x',1) + \delta v(r_2,2)) \\ &= \frac{1}{2}\left(v(x,1) + \delta\left(\frac{1}{2}v(y,2) + \frac{1}{2}v(y',2)\right)\right) + \frac{1}{2}(v(x',1) + \delta v(z,2)), \end{aligned}$$

in which δ is the individual discount factor and $v(\cdot, \tau)$ is the period-τ individual instantaneous utility function. Note that the above equation can be rewritten as

$$V_1(\tilde{p}_1) = \left(\frac{1}{2}v(x,1) + \frac{1}{2}v(x',1)\right) + \delta\left(\frac{1}{4}v(y,2) + \frac{1}{4}v(y',2) + \frac{1}{2}v(z,2)\right);$$

that is, the utility of $\tilde{p}_1 \in \Delta(X \times \Delta(X))$ is equal to the following dynamic lottery: In period

[20] For any metric space Y, let $\Delta(Y)$ denote the set of Borel probability measures on Y. We endow $\Delta(X)$ with the Prohorov metric and $X \times \Delta(X)$ with product topology.

1, the individual consumes a 50-50 lottery between x and x', and in period 2, he consumes a lottery that yields y with probability 1/4, y' with probability 1/4, and z with probability 1/2.

It is not difficult to see the logic behind this observation. In general, given any $\tilde{p}_1 \in \Delta(X \times \Delta(X))$, we compute the marginal probability distribution of period-1 outcomes and call it $p_1 \in \Delta(X)$, and compute the marginal probability distribution of period-2 outcomes and call it $p_2 \in \Delta(X)$. Then, (p_1, p_2) is a dynamic lottery whose period-2 lotteries are independent of period-1 outcomes. It must be the case that $V_1(\tilde{p}_1) = V_1((p_1, p_2))$, because V_1 is a time-additively separable expected utility function.

Second, consider the generation-2 individual. Because we are examining the period-1 planner's problem, which means the dynamic lottery's risk has not resolved, how does the planner evaluate the second generation's utility of $\tilde{p}_1$? Arguably,

$$V_2(\tilde{p}_1) = \frac{1}{2}\left(\frac{1}{2}v(y,2) + \frac{1}{2}v(y',2)\right) + \frac{1}{2}v(z,2) \tag{II.49}$$

seems to be a reasonable evaluation—with probability 1/2, the second generation's utility will be $\frac{1}{2}v(y,2) + \frac{1}{2}v(y',2)$, and with probability 1/2, the second generation's utility will be $v(z,2)$. Now, again,

$$V_2(\tilde{p}_1) = V_2((p_1, p_2)) = \frac{1}{4}v(y,2) + \frac{1}{4}v(y',2) + \frac{1}{2}v(z,2).$$

Therefore, $\tilde{p}_1$ and (p_1, p_2) are equivalent for the planner in period 1. The planner's period-1 aggregation problem under $\Delta(X \times \Delta(X))$ is the same as under $\Delta(X)^2$—there is a bijection between time-addtively separable expected utility functions defined on the domain with and without correlation. As long as the period-1 planner uses the same utilitarian weights to aggregate individual utility functions, the planner's preference will be the same in both cases.

Move on to period 2 and continue our previous example of $\tilde{p}_1$ and (p_1, p_2). With either $\Delta(X \times \Delta(X))$ or $\Delta(X)^2$, the second generation's utility function is defined on $\Delta(X)$, because individuals do not care about past consumption. Therefore, there is again a (trivial) bijection between generation-2 individual utility functions defined on the domain with and without correlation. The planner's period-2 preference will be identical in both cases as long as she uses the same utilitarian weights for individuals.

The analysis above can be extended to the case with more periods and more individuals. In this sense, focusing on consumption sequences $\Delta(X)^T$ without modeling how uncertainty resolves over time is without loss of generality.

However, it should be noted that with $\tilde{p}_1$, the period-2 lottery is either q_2 or r_2. With (p_1, p_2), no matter what the first generation consumes, the period-2 lottery is p_2. Therefore, there will be some ex post difference between $\tilde{p}_1$ and (p_1, p_2) about which generation consumes what. However, this difference should not affect the period-2 planner's aggregation problem.

Another issue to be noted is that in either the case with correlation or the case without, we only study what the planner's objective should be if she aggregates individuals' preferences. This exercise does not require us to consider, for example, feasibility constraints. If the planner's problem is to maximize some objective under certain constraints, correlation may be important in the feasibility constraints. For example, if there is a technological advancement in the first period, we can anticipate a larger feasible set of consumption in the future. This requires correlation in the constraints.

CHAPTER III

Getting Information from Your Enemies

A decision maker DM faces a binary choice. DM does not know which alternative is better, but a group of experts do. However, the experts would like DM to make the wrong choice. Given the opposing preferences, is it still possible for DM to extract useful information from the experts using mechanism design? We answer "Yes": There are mechanisms where truth-telling is a Bayesian or even ex post equilibrium, even though the information leak benefits DM and hurts the expert. On the other hand, if truth-telling is required to be an interim or ex post dominant strategy, then no mechanism extracts information in favor of DM. We also discuss two extensions. In the first extension, we show that DM can extract information even if his commitment to a mechanism is limited. In the second extension, we show that if DM can first Blackwell-garble the experts' information, then information extraction becomes much easier.

III.1 Introduction

Can someone, using neither force nor lure, and not even a trick, get his enemies to tell him secrets that benefit him and harm them? Yes, he (sometimes) can.

Suppose he is a decision maker DM who faces a binary choice: S or R? S is the Safe option — the value of it is always 0. R is Risky — its value depends on the state of the world, s. DM does not know what the state is, but a group of experts do. In particular, each expert has a piece of *partial* information about the state — s manifests once these pieces of partial information are combined.

DM wishes to collect these information fragments from the experts so he can learn about the state and make an informed choice. There is, however, a hurdle: The experts would like him to make the wrong choice, because it happens that, in every state, the lower DM's payoff, the higher the experts' payoffs. In the face of such hostility, can DM still extract

This chapter is based on the working paper "Getting Information from Your Enemies "(Feng and Wu, 2019).

information from the experts so that he makes a better choice than if he was to choose without any information? This is the question we address in the chapter.

We suppose DM can commit to a mechanism.[1] Moreover, to make the analysis interesting, suppose DM cannot make any transfers. In other words, he cannot make monetary rewards and punishments part of the mechanism, for it is clear that these tools would make information extraction easier.[2]

The situation does not look promising: The experts are hostile, and DM cannot use transfers to alleviate the severe conflict of interests in between. Rough intuition could suggest that the experts would not willingly leak information, knowing that the leak would benefit DM and harm themselves. However, surprisingly, we find that DM *can* extract information from his enemies in many cases.

More specifically, we show that there are (direct revelation) mechanisms such that truth-telling is Bayesian, or even ex post, equilibrium strategy for the experts, despite that the leaked information help DM make a better choice that in turn harms the experts. On the other hand, the possibility of such information extraction depends on the environment, which technically is captured by the payoff structure and information structure. We characterize all environments in which information extraction is possible via ex post incentive compatible mechanisms. Furthermore, we show that the weaker Bayesian incentive compatibility allows more information extraction through examples. In particular, for environments corresponding to the classical common value voting or "Condorcet Jury" models, Bayesian-incentive compatible information leak can be so serious that DM is able to almost achieve first-best when there are many experts.

On a different note, we also observe that if truth-telling has to be an *interim dominant strategy*[3] for the experts, i.e. a best response given one's private information against *any* strategies from the other experts, then information extraction is impossible. This implies that information extraction is also impossible under the conventional dominant strategy incentive compatibility, which is a stronger condition.

We consider two extensions. First, we ask whether DM can extract information without full commitment to a mechanism, and find that he indeed can, as long as he commits to a game protocol but not necessarily to a choice rule. In particular, for any direct revelation mechanism, there is a cheap talk game where the DM achieves the same equilibrium payoff

[1] It turns out, as we will show, that full commitment to a mechanism, that is, commitment to a game form *and* a choice rule, is more than necessary. The answer to the research question remains the same as long as there is commitment to the game form only.

[2] Obviously, either ex post transfers, or side bets à la Crémer and McLean (1988), can easily be used in DM's favor.

[3] The concept of interim dominant strategy is proposed and elaborately analyzed in Feng and Wu (2020) as a weakening of the dominant strategy in the interdependent value setting.

as he does using the mechanism.

In the second extension, we give DM the ability to Blackwell-garble the experts' information source, though he observes neither the original nor the garbled information. We show that, with this additional device, DM can translate an environment in which information extraction is not possible into one in which it is, thus making information extraction easier.

In the chapter, we try to understand whether mechanism design helps at all if there is a severe conflict of interests between the uninformed designer and the informed agents.

In most mechanism design and contract theoretic models, despite some misalignment of interests between the designer and the agents, there typically is a common ground where everyone benefits from some information sharing. This potential surplus can easily be (mis)understood as a foundation for meaningful mechanism design, as facilitating information sharing is what a mechanism does. Intuitively, if sharing information would hurt, why would the agents willingly reveal information even if participation is compulsory?

In this chapter we challenge, and then subvert, this intuition, by showing that a mechanism helps even if the informed parties are enemies to the designer. This finding is a significant contribution to the literature, as it establishes that a potential surplus is not necessary for information sharing.

The chapter can also be viewed as investigating how dramatically may a collective choice institution fail. Institutions like voting systems serve the purpose of aggregating information distributed among the population to arrive at a collective choice. We show in this chapter that incentive compatible actions from the participants may lead to a collective choice that is unsatisfactory to everyone. Indeed, the collective choice based on aggregated information can in expectation be worse than a naive coin toss. Given the finding, there are situations in which, when designing a voting system, we have to caution against potential failure that may generate outcomes opposite to the public interest.

Related Literature

The overarching theme of this papechapterr is an important question: When is information sharing possible? There has been extensive research on this theme, and a common finding is that information sharing tends to break down when the conflict of interests between the participating parties becomes severe. For example, in the standard cheap talk model of Crawford and Sobel (1982), information sharing is not possible when the preference of the single informed expert is sufficiently different from that of the uninformed DM. On the other hand, when there are more experts, then a high degree of correlation between the experts' information can help DM extract information. A stream of papers, including Gilligan and

Krehbiel (1989), Krishna and Morgan (2001), Battaglini (2002, 2004), Gerardi, McLean and Postlewaite (2009), Ambrus and Lu (2014), are based on this idea.

Without assuming highly correlated private information, Wolinsky (2002) shows that when DM faces a group of partially informed experts, information sharing is possible despite the experts' significantly biased preferences relative to DM's. The setup of Wolinsky (2002), particularly in that there are multiple partially informed experts who share similar preferences, is similar to ours, and moreover the results of his paper and ours are also in the same spirit. However, the conflict of interests between DM and the experts is much more severe in our model than in Wolinsky (2002) — their preferences are always diametrically opposed in our model. Therefore our results showing the possibility of information sharing is stronger.

Organization

The chapter is organized as follows: Section 2 describes the model. Section 3 delivers the analysis and the results. Section 4 concludes. All proofs are in the Appendix.

III.2 Model

This section describes the model.

Environment

A decision maker DM needs to make a choice between two options, S and R, which affects himself and a group of N experts (indexed as expert 1, ..., expert N) . If option S is chosen, every player including DM receives a payoff of 0. If option R is chosen, the payoffs depend on an $N-$dimensional state of the world $(s_1, ..., s_N) \in S_1 \times ... \times S_N = S$. In particular, the payoff from choosing R to each expert is $e(s)$ and the payoff to DM is $-e(s)$. Observe that the interests between DM and the experts are diametrically opposed at both the ex ante and the ex post stage — this makes them enemies.

For simplicity assume S is finite. The probability that state $s \in S$ obtains is $p(s)$. And $p(s)$ is the common prior. Let S_p denote the set of all states s where $p(s) > 0$. Moreover denote S_+ (resp. S_-) as the set of all states s in S_p where $e(s) > 0$ (resp. $e(s) < 0$). To avoid the trivial case of DM achieving first-best without any information, assume S_+ and S_- are nonempty.

Conditional on state $s = (s_1, ..., s_n)$, expert $n = 1, ..., N$ privately observes s_n, which we will call his signal. Hence information about the state is distributed — each expert observes

only one dimension of it. DM observes nothing about s. For parsimony suppose for any i and $s_i \in S_i$ there is some $s_{-i} \in S_{-i}$ such that $p(s_i, s_{-i}) > 0$.

The environment is denoted by the environment parameters (S, e, p).

Mechanism

If DM gets no help from the experts, the best he can do is choosing the option that generates a higher ex ante payoff. DM's expected payoff from choosing this ex ante better option without using any information is

$$B := \max\{0, -\sum_{s \in S} p(s)e(s)\},$$

whereas an expert's expected payoff from this choice is $-B$. We call B the (no-information) **benchmark**. Can DM do better than the benchmark by trying to get information from the experts? By the Revelation Principle, if he can in any way then he can by using a direct revelation mechanism. Therefore we shall focus on direct revelation mechanisms. As discussed in the Introduction, we consider only mechanisms without transfers, as transfers would make information extraction much easier.

A direct revelation mechanism is given by a choice rule $q : S \to [0, 1]$, where $q(s_1, ..., s_N)$ is the probability that R is chosen given each expert $i = 1, ..., N$ reporting s_i. We say that mechanism q **extracts information** (for DM) if truth-telling is an equilibrium [4] under q, and moreover DM's expected payoff from the mechanism is greater than the benchmark B.

Solution concepts

In this chapter we analyze information extraction under the following four kinds of mechanisms, each of which associated with a different set of incentive compatibility constraints:

- Dominant strategy incentive compatible mechanisms
- Ex post incentive compatible mechanisms
- Interim dominant strategy incentive compatible mechanisms
- Bayesian incentive compatible mechanisms.

Dominant strategy incentive compatibility requires that for every expert, truth-telling is a best response given *any* beliefs about the following:

(1) The distribution of the other experts' signals

[4] We shall analyze implications of different equilibrium concepts including dominant strategy equilibrium, ex post equilibrium, interim dominant strategy equilibrium, and Bayes Nash equilibrium.

(2) The other experts' strategies.

Ex post incentive compatibility requires truth-telling be a best response given any belief about (1) and a *correct* belief about (2) (that the other experts are also truth-telling).

Interim dominant strategy incentive compatibility, which we discuss in more detail in Feng and Wu (2020), requires truth-telling be a best response given any belief about (2) and a *correct* belief about (1) (that signals are distributed according to p).

Bayesian incentive compatibility assumes correct beliefs about both (1) and (2).

It is immediate that Bayesian incentive compatibility is the weakest and dominant strategy incentive compatibility is the strongest. Ex post incentive compatibility and interim dominant strategy incentive compatibility, on the other hand, are not mutually comparable in general.

Participation constraint

We do not impose a participation constraint. In other words, experts have to participate in the mechanism. Given our setting, DM's choice has a public good nature — it affects the experts regardless of whether they participate in the collective choice process (the mechanism) or not. Therefore, to come up with a reasonable participation constraint, we must endogenously derive the value of not participating (the outside option).

If we suppose the experts can *collectively* opt out of the mechanism, then, should that happen, DM will choose the option that gives him the benchmark payoff of B, and each expert gets a payoff of $-B$. It is thus natural to use $-B$ as the value of the outside option in the participation constraint, but this immediately implies that information extraction is impossible, since in any mechanism that extracts information an expert's payoff is less than $-B$. The problem thus becomes trivial.

Alternatively, if experts can only *individually* opt out of the mechanism, then, to derive the value of the outside option, we need to explicitly model the pre-mechanism game in which experts decide whether to participate or not. This is certainly a worthwhile exercise, yet it is beyond the scope of this chapter to examine this more complicated procedure.

In our setting, compulsory participation is not as strong an assumption as it seems. Indeed, an expert has the freedom to "remain silent" by babbling, or to tell a lie, which will not cause additional punishment because of the absence of transfers.

III.3 Analysis

III.3.1 Dominant Strategy Incentive Compatible Mechanisms

We first formally define dominant strategy incentive compatibility.

Definition. Mechanism q is **dominant strategy incentive compatible (DSIC)** if for any $s \in S_p$, $i = 1, ..., N$ and $s'_i \in S_i, s'_{-i} \in S_{-i}$,

$$e(s)q(s_i, s'_{-i}) \geq e(s)q(s'_i, s'_{-i}).$$

Thus truth-telling is a dominant strategy for every expert under q. The following lemma characterizes DSIC mechanisms.

Lemma III.1. *q is DSIC if and only if for any $i = 1, ..., N$ and $s_i \in S_i$:*

1. *If $(s_i, s_{-i}) \in S_+$ for some $s_{-i} \in S_{-i}$, then $q(s_i, s'_{-i}) \geq q(s'_i, s'_{-i})$ for any $s'_i \in S_i$ and $s'_{-i} \in S_{-i}$.*
2. *If $(s_i, s_{-i}) \in S_-$ for some $s_{-i} \in S_{-i}$, then $q(s_i, s'_{-i}) \leq q(s'_i, s'_{-i})$ for any $s'_i \in S_i$ and $s'_{-i} \in S_{-i}$.*

The following lemma states that a DSIC mechanism must take a particular form that assumes at most three values.

Lemma III.2. *If q is DSIC then there exist $q_+, q_- \in [0, 1]$ where $q_+ \geq q_-$, such that $q(s) = q_+$ for every $s \in S_+$, $q(s) = q_-$ for every $s \in S_-$, and $q(s) \in [q_-, q_+]$ for every $s \notin S_p$.*

The following lemma states that if every expert i may get some signal s_i that is not perfectly informative of whether R or S is the better option for the experts, then a DSIC mechanism must be a constant mechanism.

Lemma III.3. *If for every $i = 1, ..., N$ there exist $s_i \in S_i$ and $s_{-i}, s'_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in S_+$ and $(s_i, s_{-i}) \in S_-$, then q is DSIC if and only if it is a constant mechanism.*

What kind of environment does Lemma III.3 leave out? Not much. If the condition in the lemma is not satisfied, then there must be an expert i such that given any $s_i \in S_i$, either $e(s_i, s_{-i}) > 0$ for all $s_{-i} \in S_{-i}$ where $(s_i, s_{-i}) \in S_p$ or $e(s_i, s_{-i}) < 0$ for all $s_{-i} \in S_{-i}$ where $(s_i, s_{-i}) \in S_p$. In other words, i's signal alone is perfectly informative of whether R or S is

the better option for the experts. We call i the **informed expert**[5], and denote S_i^R (resp. S_i^L) as the set of signals that are perfectly informative of R (resp. S) being the better option for the experts. Hence Lemma III.3 can be restated as the following: A DSIC mechanism is a constant mechanism if there is no informed expert. On the other hand, the following lemma characterizes all dominant strategy mechanisms when there exist informed experts.

Lemma III.4. *If there exists an informed expert i then q is DSIC if and only if there are $q_+, q_- \in [0,1]$ where $q_+ \geq q_-$, such that $q(s) = q_+$ if $s_i \in S_i^R$ and $q(s) = q_-$ if $s_i \in S_i^L$.*

The proof is skipped, as the lemma follows straightforwardly from Lemmas III.1 and III.2.

It follows from Lemmas III.3 and III.4 that DM cannot extract information using a DSIC mechanism, because either there is no informed expert so the mechanism is constant, which is weakly worse than what DM can get without any information, or there is an informed expert so the mechanism has to cater to his preference, which is in direct conflict to DM's.

Proposition III.1. *There is no DSIC mechanism that extracts information.*

III.3.2 Ex Post Incentive Compatible Mechanisms

Ex post incentive compatibility is formally defined as follows:

Definition. Mechanism q is **ex post incentive compatible (EPIC)** if for any $s \in S_p$, $i = 1, ..., N$ and $s_i' \in S_i$,

$$e(s)q(s) \geq e(s)q(s_i', s_{-i}).$$

Therefore, under an EPIC mechanism, truth-telling is a best response for every expert even if the state is common knowledge. It equivalently means that truth-telling is a best response regardless of an expert's belief about the distribution of the other experts' signals. In contrast with DSIC, EPIC assumes every expert to (correctly) believe that the other experts are truthful.

We say that s and s' are **contiguous** if they differ in only one entry, or they are **close** if they differ in one or two entries.

The following lemma, which "almost" characterizes EPIC mechanisms, will be useful.

Lemma III.5. *q is EPIC only if for any $s, s' \in S$:*

[5]Note that an informed expert is not perfectly informed of all payoff-relevant information, because he may still need the other experts' information to fully know the exact value of $e(s)$.

1. *$q(s) = q(s')$ if s and s' are contiguous, and s, s' are both in S_+ or both in S_-.*

2. *$q(s) \geq q(s')$ if s and s' are close, and $s \in S_+$ and $s' \in S_-$.*

Moreover for any q satisfying Conditions 1 and 2 there exists an EPIC mechanism $\tilde{q}$ such that $\tilde{q}(s) = q(s)$ for every $s \in S_p$.

To prepare for a characterization of all environments in which there is an EPIC mechanism that extracts information, introduce some additional terminology. For any $C \subset S_p$ let $v(C)$ denote $-\sum_{s \in C} p(s)e(s)$. Define $C \subset S_p$ as a **cluster** if:

C1: No $s \in C$ is contiguous to any $s' \in S_p - C$ where $\text{sgn}(e(s')) = \text{sgn}(e(s))$.

C2: One of the following is true:

(A) $v(C) > 0$ and no $s \in C \cap S_-$ is close to any $s' \in (S_p - C) \cap S_+$.

(B) $v(C) < 0$ and no $s \in C \cap S_+$ is close to any $s' \in (S_p - C) \cap S_-$.

If C satisfies Conditions C1 and C2(A) (resp. C2(B)) then we call it a **Type A** (resp. **Type B**) **cluster**.

Proposition III.2. *There exists an EPIC mechanism that extracts information if and only if the environment (S, e, p) satisfies one of the following is true:*

1. *$v(S_p) \leq 0$ and there is a Type A cluster.*

2. *$v(S_p) \geq 0$ and there is a Type B cluster.*

The following proposition characterizes a special class of environments in which DM can use an EPIC mechanism to achieve a payoff that is not only higher than the benchmark, but actually reaches the **first-best** level, which is the payoff DM would get if he knew the state.

An environment (S, e, p) is said to be **rich** if there exist two states s and s' satisfying:

- $p(s) > 0$ and $p(s') > 0$
- $\text{sgn}(e(s)) \neq \text{sgn}(e(s'))$
- s and s' differ in the signals of no more than two experts.

Proposition III.3. *DM can achieve his first-best payoff using an EPIC mechanism if and only if the environment is not rich.*

Remark 1: Proposition III.3 is probably better read as an impossibility result. The condition stated in the proposition is difficult to satisfy. If viewing the state space as a lattice in which states with common entries are connected, an environment satisfying the condition must have enough zero-probability states that separate positive-probability state-clusters where $e(s) > 0$ from negative-probability state-clusters where $e(s) < 0$. For example, if $N \leq 2$ then no environment satisfies the condition.

It is well known that if $N > 2$ and experts' signals are perfectly correlated, then full information extraction is straightforward. Proposition III.3 is a generalization of this result, as such an information structure can be easily verified to satisfy the condition in Proposition III.3.

Remark 2: The proof shows more than stated in the proposition. Indeed, by the proof, it is impossible for DM to get the first-best payoff with a mechanism that is just Bayesian incentive compatible but not EPIC. Hence, as long as Bayesian incentive compatibility needs to be satisfied, any mechanism that gives DM the first-best payoff must be EPIC.

III.3.3 Interim Dominant Strategy Incentive Compatible Mechanisms

Interim dominant strategy incentive compatibility is a condition we analyze in depth, for a much more general setting, in Feng and Wu (2020). The formal definition is given as follows:

Definition. Mechanism q is **interim dominant strategy incentive compatible (IDSIC)** if for any $i = 1, ..., N$, $s_i, s_i' \in S_i$, and any function $z : S_{-i} \to S_{-i}$,

$$\begin{aligned} &\sum_{s_{-i} \in S_i} Pr(s_{-i}|s_i) e(s_i, s_{-i}) q(s_i, z(s_{-i})) \\ \geq &\sum_{s_{-i} \in S_i} Pr(s_{-i}|s_i) e(s_i, s_{-i}) q(s_i', z(s_{-i})), \end{aligned}$$

where $Pr(s_{-i}|s_i)$ is derived from the common prior $p(s)$.

Under an IDSIC mechanism, truth-telling is a best response as long as an expert has the correct belief about the distribution of the other experts' signals, regardless of what he believes about the other experts' strategies.[6]

[6] Although the definition is given in terms of truth-telling as a best response to any possible pure strategy profile from the other players, it immediately implies that truth-telling is a best response to any possible mixed strategy profile from the other players as well.

Denote $h(s) := p(s)e(s)$. For any expert i and $s_i \in S_i$ define

$$\underline{\alpha}_i(s_i) := \sum_{s_{-i} \in S_{-i}}^{h(s_i, s_{-i}) < 0} h(s_i, s_{-i}),$$

$$\overline{\alpha}_i(s_i) := \sum_{s_{-i} \in S_{-i}}^{h(s_i, s_{-i}) \geq 0} h(s_i, s_{-i}),$$

and

$$\beta_i(s_i) := \sum_{s_{-i} \in S_{-i}} h(s_i, s_{-i}).$$

Clearly we have $\underline{\alpha}_i(s_i) \leq 0 \leq \overline{\alpha}_i(s_i)$ and $\underline{\alpha}_i(s_i) + \overline{\alpha}_i(s_i) = \beta_i(s_i)$.

The following lemma characterizes all IDSIC mechanisms.

Lemma III.6. *q is IDSIC if and only if for any $i = 1, ..., N$, $s_i, s_i' \in S_i$, and $s_{-i}, s_{-i}' \in S_{-i}$:*

$$\underline{\alpha}_i(s_i)\Big(q(s_i, s_{-i}) - q(s_i', s_{-i})\Big) + \overline{\alpha}_i(s_i)\Big(q(s_i, s_{-i}') - q(s_i', s_{-i}')\Big) \geq 0.$$

Based on the lemma, we show that it is impossible to extract information with an IDSIC mechanism.

Proposition III.4. *There is no IDSIC mechanism that extracts information.*

Remark: Since every DSIC mechanism is also an IDSIC mechanism, Proposition III.1 can also be derived immediately as a corollary of Proposition III.4.

III.3.4 Bayesian Incentive Compatible Mechanisms

Bayesian incentive compatibility is formally defined as follows:

Definition. Mechanism q is **Bayesian incentive compatible (BIC)** if for any $i = 1, ..., N$ and $s_i, s_i' \in S_i$,

$$\begin{aligned} &\sum_{s_{-i} \in S_{-i}} Pr(s_{-i}|s_i) e(s_i, s_{-i}) q(s_i, s_{-i}) \\ \geq &\sum_{s_{-i} \in S_{-i}} Pr(s_{-i}|s_i) e(s_i, s_{-i}) q(s_i', s_{-i}), \end{aligned}$$

where $Pr(s_{-i}|s_i)$ is derived from the common prior $p(s)$.

Clearly, BIC is implied by IDSIC, EPIC or DSIC.

The following proposition shows that two environments (S, e, p) and (S, e', p') are equivalent in terms of information extraction using BIC mechanism, if $p(s)e(s) = p'(s)e'(s)$ for any $s \in S$.

Proposition III.5. *Fix two environments (S, e, p) and (S, e', p') where $p(s)e(s) = p'(s)e'(s)$ for every $s \in S$. If mechanism q is a BIC mechanism that extracts information in (S, e, p), then it is a BIC mechanism that extracts information in (S, e', p').*

Remark. An immediate implication of Proposition III.5 is that in a full-support environment, the possibility of information extraction does not crucially depend on how signals are correlated, because if there is a BIC mechanism q that extracts information where p has full support, then for any full-support p' we can find payoff function e' such that $p(s)e(s) = p'(s)e'(s)$, which by Proposition III.5 implies q is also a BIC mechanism that extracts information in (S, e', p'). This result is in contrast with results from a number of papers which rely on correlation between the experts' signals for information sharing.[7]

Since EPIC implies BIC, from Proposition III.2 we immediately know that there exist environments in which DM can extract information using a BIC mechanism. Can DM do more as EPIC is weakened to BIC? Yes, indeed. Below we show two examples in which an EPIC mechanism cannot extract information where a BIC mechanism can.

Example 1

There are two experts, who each can observe a signal of either 0 or 1. The payoffs and probabilities are given as follows.

e

	0	1
0	-4	2
1	2	-1

p

	0	1
0	1/4	1/4
1	1/4	1/4

Consider the following mechanism:

q

	0	1
0	1	2/3
1	2/3	0

[7] See the Introduction for a discussion on this literature.

DM's expected payoff from the mechanism is 1/3, which is higher than the benchmark of 1/4.

Observe that this environment does not allow information extraction in EPIC mechanism, because there are only two experts.

Example 2

There are N experts, who each can observe a signal of either 0 or 1. If at least n experts observe signal 1 then the value of R is 1. Otherwise the value of R is $-1 < 0$. All states (signal profiles) are equally likely.

This example can be interpreted as a variation of the Condorcet Jury problem: S correspond to acquittal, and R correspond to conviction. If at least n experts observe 1 then the defendant is guilty (in which case the experts, who are the jurors, would like a conviction). Otherwise the defendant is innocent.

For simplicity suppose $n \geq N/2$ so that $B = 0$.

Consider the following mechanism:

- $q(s) = 1$ if $\sum_{i=1}^{N} s_i \leq n - 2$.
- $q(s) = 0$ if $\sum_{i=1}^{N} s_i \geq n + 1$.
- $q(s) = \frac{\binom{N-1}{n-2}}{\binom{N-1}{n}+\binom{N-1}{n-2}}$ if $\sum_{i=1}^{N} s_i = n - 1$ or n.

It is straightforward to verify that q is a BIC mechanism that extracts information as long as $N \geq 3$ and gives the designer an expected payoff arbitrarily close to the first-best payoff when N goes to infinity.

The rest of this chapter focuses on Condorcet Jury votings.

Generalized Condorcet Jury

Example 2 corresponds to an important class of collective choice problems known as "Cordorcet Jury Problems". We define a generalized Condorcet Jury problem as the following. N agents each gets a binary signal taking value of 0 or 1. $e(s)$ is permutation invariant, so the value of R depends on how many 1-signals obtain. $p(s)$ is permutation invariant. Let $\epsilon(s) := \sum_{i=1}^{N} s_i$, $e(s) \geq e(s')$ if and only if $\epsilon(s) \geq \epsilon(s')$, that is, R is more valuable if there are more 1-signals. This situation generalizes the Condorcet Jury model with the interpretation that the agents are jurors to determine whether to convict or acquit a defendant. A 1-signal is a partial evidence that the defendant is guilty, thus the more 1's the more likely the defendant is guilty. S represents the decision to acquit and R to convict. The jurors prefer to acquit if there's no enough guilty evidence or to convict otherwise.

DM's optimization problem is the following:

$$\max_{0\leq q(s)\leq 1} -\sum_{s\in S} p(s)e(s)q(s)$$

$$\text{s.t.} \sum_{s_{-i}\in S_{-i}} p(s_i,s_{-i})e(s_i,s_{-i})q(s_i,s_{-i}) \geq \sum_{s_{-i}\in S_{-i}} p(s_i,s_{-i})e(s_i,s_{-i})q(s_i',s_{-i})$$
$$\text{for all } i=1,...,N,\ s_i\in S_i \text{ and } s_i'\in S_i.$$

For any given $s=(s_1,s_2,\cdots,s_n)$, there all $n!$ permutations of s in total. We sort all permutations in lexicographic order, i.e.,

$$\sigma_1(s)=(s_1,s_2,\cdots,s_n),\sigma_2(s)=(s_1,s_2,\cdots,s_n,s_{n-1}),\cdots,\sigma_{n!}(s)=(s_n,s_{n-1},\cdots,s_1).$$

Lemma III.7. *Suppose $p(s)$ and $e(s)$ are permutation invariant. If $q(s)$ is a solution to the above problem, then*

$$q^*(s)=\frac{1}{n!}\sum_{k=1}^{n!} q(\sigma_k(s))$$

is a solution.

Given Lemma III.7 we can focus on symmetric mechanisms. A symmetric mechanism can be expressed as $q_{sym}:\{0,...,N\}\to[0,1]$ where $q_{sym}(k)$ is the probability that R is chosen given a report that has k 1-signals. Define $e_k=e(s)$, $p_k=p(s)$ and $h_k=e_kp_k$ where $\epsilon(s)=k$.

We first take a detour and ask an unusual question whether there is a BIC mechanism that has a *decreasing* rules, that is, the probability of conviction goes *down* as evidence for guilty becomes strong. Here we do not consider the trivial cases of constant mechanisms.

Lemma III.8. *There exists a decreasing BIC mechanism if and only if there are $i,j\in\{0,...,N\}$ where $h_i,h_{i+1}<0$, $h_j,h_{j+1}>0$, and $\frac{h_{i+1}}{h_i}\geq\frac{h_{j+1}}{h_j}$.*

Lemma III.8 characterizes all environments that admit a BIC decreasing mechanism. Combined with Lemma III.8, the following Lemma provides a sufficient condition for DM to beat the benchmark.

Lemma III.9. *If the DM is indifferent between S and R ex ante, then any decreasing mechanism beats the benchmark.*

Lemma III.9 shows in any environment that the DM is indifferent between S and R ex ante ($\sum_{k=0}^{N}\binom{N}{k}h_k = 0$), the existence of a BIC decreasing mechanism is a sufficient condition for beating the benchmark. However, existence of a BIC decreasing mechanism is not a necessary condition for beating the benchmark incentive compatibly. Consider the environment where $N = 3$, $h_0 = -2$, $h_1 = -1$, $h_2 = 1$, $h_3 = 2$. Lemma III.8 implies that a BIC decreasing mechanism does not exist. However, the mechanism where $q_0 = 1$, $q_1 = 1/4$, $q_2 = 3/4$, $q_3 = 0$ is BIC and beats the benchmark.

Another question is whether, when $\sum_{k=0}^{N}\binom{N}{k}h_k$ is not necessarily 0, the best decreasing BIC mechanism would still beat the benchmark. This answer to this question is again negative. Consider the environment where $N = 3$, $h_0 = h_1 = -1$, $h_2 = h_3 = 10$. Then all decreasing mechanisms are in the form of $q_0 = x, q_1 = y, q_2 = y, q_3 = z$ where $\frac{x-y}{y-z} = 10$. The expected utility for the agents from this mechanism is

$$-x - 3y + 30y + 10z = 16y + 20z \geq 0,$$

Implying no decreasing BIC mechanism beats the benchmark.

Classical Condorcet Jury

This part considers the classical Condorcet Jury problem. The ex ante probability of the defendant being guilty is π. A juror's utility from convicting a guilty defendant is x and that from convicting an innocent defendant is $-y$, where x and y are positive. Conditional on the defendant being guilty, every juror independently receives the guilty (1) signal with probability $\alpha > 1/2$. Conditional on the defendant being innocent, every juror independently receives the guilty (1) signal with probability $\beta < 1/2$.

Proposition III.6. *In the classical Condorcet model, there is no symmetric monotone BIC mechanism that extracts information.*

It is straightforward to verify that h_k is increasing in k in the classical Condorcet model. Hence, by Lemma III.8 there does not exist a decreasing BIC mechanism for the classical Condorcet Jury model. And it is easy to see that no symmetric increasing BIC mechanism that extracts information. Thus no symmetric monotone BIC mechanism that extracts information. By Lemma III.7 we know there is no monotone BIC mechanism that extracts information.

As mentioned in the introduction, this chapter can also be viewed as investigating how dramatically may a collective choice institution fail. Institutions like voting systems serve the

purpose of aggregating information distributed among the population to arrive at a collective choice.

As shown in example 2, incentive compatible actions from the participants may lead to a collective choice that is unsatisfactory to everyone. Given the finding, there are situations in which, when designing a voting system, we have to caution against potential failure that may generate outcomes opposite to the public interest. The next proposition brings a reassuring message that the voting rule being used everyday never harms voters in the classical Condorcet model.

Proposition III.7. *Suppose in the classical Condorcet model, the voting rule is:*

1. *symmetric with respect to jurors,*
2. *monotone in the number of guilty reports.*

Then any equilibrium payoff of the jurors is weakly higher than benchmark payoff $-B$.

It is well known that given a voting rule, there may exist multiple equilibria. Proposition III.7 tells us a symmetric and monotone voting rule gives a higher expected payoff to the jurors under any equilibrium compared with the jurors' no-voting benchmark, $-B$ (see page 144). Hence, a symmetric and monotone voting rule never fails in the classical Condorcet model.

To understand Proposition III.7, consider a voting rule where the verdict is conviction if there are at least $\overline{k}$ reports of the guilty (1) signal, or acquittal otherwise. Suppose jurors follow the same strategy of reporting 1 with probability a conditional on the guilty signal or with probability b conditional on the innocent signal. Compare two states s and s' that differ only in juror i's signal, where $s_i = 0$ and $s'_i = 1$. Let Q_s and $Q_{s'}$ respectively denote the probability of conviction in states s and s'. Let $\tilde{p}_k$ denote the probability that the non-i jurors report k guilty signals in total conditional on s_{-i}. Thus

$$\begin{aligned} Q_{s'} - Q_s =& \Big[\sum_{k=\overline{k}}^{N-1} \tilde{p}_k + \tilde{p}_{\overline{k}-1} a\Big] - \Big[\sum_{k=\overline{k}}^{N-1} \tilde{p}_k + \tilde{p}_{\overline{k}-1} b\Big] \\ =& \tilde{p}_{\overline{k}-1}(a-b). \end{aligned}$$

By symmetry of the strategy profile, the implied probability of conviction conditional on any state only depends on the number of actual guilty signals that obtain in the state. Let $\tilde{q}_k$ denote this probability if the number of guilty signals that obtain is k. The above inequality implies $\tilde{q}_k$ is monotone in k. By the Revelation Principle, a symmetric strategy profile (a, b) is an equilibrium only if the implied (monotone) direct mechanism $(\tilde{q}_k)$ is BIC.

Since no monotone BIC mechanism extracts information, we conclude that no symmetric equilibrium of the voting game makes the jurors worse off than no voting. The proof shows that for any (possibly non-symmetric) equilibrium the implied direct BIC mechanism is monotone. Hence, any equilibrium payoff of the jurors is weakly higher than the no-voting benchmark.

III.4 Extensions

III.4.1 Extension 1: without Commitment

There is a shortcoming to the direct revelation mechanism: It requires full commitment from DM, which makes the mechanism impractical in some situations. Indeed, if DM and the experts are enemies it would be difficult to convince the experts that DM would honor the choice rule q. Thus it is worth asking whether information extraction is still possible if some of the commitment requirement is relaxed.

By using a mechanism, DM commits to two things: 1. A game form, and 2. A choice rule given the game outcome. Below we show that as long as DM can commit to a game form (communication protocol), commitment to a choice rule is not necessary for information extraction. In particular, as long as there is a BIC direct revelation mechanism q that extracts information, the same payoff can be achieved by cheap-talk communication. More specifically, given direct revelation mechanism q let $\Gamma(q)$ be the associated cheap talk game defined in the following sense:

Step 1 Experts send messages simultaneously, where the message set for expert i is S_i.

Step 2 Given message profile s, a public randomization device generates message R with probability $q(s)$ and message S with probability $1 - q(s)$.

Step 3 DM makes choice after observing the device's recommendation. He does not observe messages from the experts.

The following result shows that there is a perfect Bayesian equilibrium of the cheap talk game which implements the choice rule q.

Proposition III.8. *$\Gamma(q)$ has a perfect Bayesian equilibrium that is outcome-equivalent to the truth-telling equilibrium under q if q is a BIC mechanism that extracts information.*

Conditional on DM follows the device's recommendation, truthtelling is incentive compatible for experts under q immediately implies truthtelling is optimal under $\Gamma(q)$. Conditional on every expert sends the message that is the same as his signal, mechanism q extracts information implies following recommendation is optimal.

III.4.2 Extension 2: Information Manipulation

In this extension we suppose DM can contaminates the experts' information, and examine whether that makes information extraction easier. Suppose that, although DM has no direct access to the experts' information, he can manipulate the information that the experts receive, that is, he may add noise to the experts' information source and distort what the experts observe. Information manipulation takes the form of garbling a là Blackwell: As DM contaminates the information source, if state s obtains, instead of receiving signal s_i with certainty, expert i receives some $s_i' \in S_i'$ with some probability, where S_i' is the set of signals that i can receive after the contamination. DM still does not observe the experts' signals, before or after contamination.

Formally, an **information manipulation** is represented by distorted state space $S' = S_1' \times ... \times S_N'$ and **garbling probability functions** $z(\cdot|s), s \in S$ on S' where $z(s'|s)$ is the probability that distorted state s' obtains, in which case expert i observes signal s_i' instead of s_i.

The probability of distorted state s' is $p'(s') := \sum_{s\in S} p(s)z(s'|s)$. Conditional on distorted state $s' \in S'$, the (expected) value of option S remains 0 for DM and the experts, and the value of option R to the experts is equal to $e'(s') := \sum_{s\in S} \Pr(s|s')e(s) = \sum_{s\in S} \frac{p(s)z(s'|s)}{p'(s')}e(s)$, and similarly the value of option R to DM is equal to $-\sum_{s\in S} \frac{p(s)z(s'|s)}{p'(s')}e(s)$.

Observe that

$$\sum_{s'\in S'} p'(s')e'(s') = \sum_{s'\in S'}\sum_{s\in S} p(s)z(s'|s)e(s) = \sum_{s\in S}\sum_{s'\in S'} p(s)z(s'|s)e(s) = \sum_{s\in S} p(s)e(s),$$

which implies that $\max\{-\sum_{s'\in S'} p'(s')e'(s'), 0\} = \max\{-\sum_{s\in S} p(s)e(s), 0\}$. In other words, the maximum payoff that DM can get without consulting with the experts remains the same regardless of information manipulation. Therefore, information manipulation might be helpful only if DM can extract some information from the experts.

For better exposition of the coming results for any state space S we (arbitrarily) order states as $s(1), ..., s(|S|)$, and given environment (S, e, p) denote $\mathbf{h} := (p(s(j))e(s(j)))_{j=1}^{|S|}$ and $\mathbf{g} := (-p(s(j))e(s(j)))_{j=1}^{|S|}$.

Lemma III.10. *DM can translate environment (S, e, p) into an environment equivalent (in the sense of Proposition III.5) to (S', e', p') by information manipulation if and only if there is a $|S'| \times |S|$ matrix Z such that:*

1. *Z is non-negative.*
2. *$Z\mathbf{h} = \mathbf{h}'$ and $Z\mathbf{g} = \mathbf{g}'$.*
3. *$Z^T \mathbf{1}_{|S'|} = \mathbf{1}_{|S|}$ where $\mathbf{1}_d$ denotes a vector of dimension d whose entries are 1's.*

Following the lemma, we can completely determine when one environment can be manipulated into another.

Proposition III.9. *DM can translate environment (S, e, p) into an environment equivalent to (S', e', p') (in the sense of Proposition III.5) by information manipulation if and only if*

1. $$\sum_{s \in S} h(s) = \sum_{s' \in S'} h'(s').$$
2. $$\sum_{s \in S}^{h(s)>0} h(s) \geq \sum_{s' \in S'}^{h'(s)>0} h'(s').$$

The proposition implies that information manipulation is very powerful. Indeed, as long as there are three experts, DM can always get his first-best payoff by translating the original environment (S, e, p) into (S', e', p') that satisfies the condition in Proposition III.3 where $S' = \{0, 1\}^N$, $h'(0, ..., 0) = \sum_{s \in S_-} h(s)$, $h'(1, ..., 1) = \sum_{s \in S_+} h(S)$ and $p'(s) = 0$ for any $s \notin \{(0, ..., 0), (1, ..., 1)\}$.

III.5 Conclusion

In this chapter we ask when a decision maker can use a mechanism without transfers to extract information from a group of experts whose preferences are diametrically opposed to his. We separately characterize environments in which this can be done subject to (1) dominant strategy incentive compatibility, (2) ex post incentive compatibility, (3) interim dominant strategy incentive compatibility, (4) Bayesian incentive compatibility. In particular, we show that in cases (1) and (3) the decision maker cannot extract information at all, whereas in cases (2) and (4) he can. We also explore the role of information manipulation and commitment in information extraction.

III.6 Proof

III.6.1 Proof of Lemma III.1

Proof. Suppose q satisfies conditions 1 and 2 in the statement. Fix $s \in S_p$ where $e(s) > 0$. It follows immediately from condition 1 that $e(s)q(s_i, s'_{-i}) \geq e(s)q(s'_i, s'_{-i})$ for any $s'_i \in S_i$ and $s'_{-i} \in S_{-i}$. The case where $e(s) < 0$ is symmetric.

Now prove the "only if" direction. Suppose there is some $s \in S_p$ such that $e(s) > 0$ and $q(s_i, s'_{-i}) < q(s'_i, s'_{-i})$ for some $s'_i \in S_i, s'_{-i} \in S_{-i}$. Thus we have $e(s)q(s_i, s'_{-i}) < e(s)q(s'_i, s'_{-i})$, implying that q is not a dominant strategy mechanism. A violation of condition 2 leads to the same conclusion. □

III.6.2 Proof of Lemma III.2

Proof. Suppose q is a dominant strategy mechanism. Pick any $s \in S_+$. Let s^k denote a state where its first k entries agree with the first k entries of s' and its last $N-k$ entries agree with the last $N-k$ entries of s. Clearly $s^0 = s$ and $s^N = s'$. Fix any $k = 1, ..., N$. Observe that $q(s^{k-1}) = q(s_k, s^{k-1}_{-k}) = q(s_k, s^k_{-k}) \geq q(s'_k, s^k_{-k}) = q(s^k)$ where the inequality is due to Lemma III.1. Thus $q(s) \geq q(s')$. If $s' \in S_+$ then likewise we have $q(s') \geq q(s)$, implying $q(s) = q(s')$. If $s, s' \in S_-$ then a symmetric argument applies, leading to $q(s) = q(s')$ again. □

III.6.3 Proof of Lemma III.3

Proof. Suppose for every $i = 1, ..., N$ there exist $s_i \in S_i$ and $s_{-i}, s'_{-i} \in S_{-i}$ such that $(s_i, s_{-i}) \in S_+$ and $(s_i, s'_{-i}) \in S_-$. Any constant mechanism is obviously a dominant strategy mechanism. To show the "only if" direction, suppose q is a dominant strategy mechanism. Fix any $i = 1, ..., N$. Since $(s_i, s_{-i}) \in S_+$, it follows from Lemma III.1 that $q(s_i, \hat{s}_{-i}) = \max_{\sigma_i \in S_i} q(\sigma_i, \hat{s}_{-i})$ for any $\hat{s}_{-i} \in S_i$. Similarly, $q(s_i, \hat{s}_{-i}) = \min_{\sigma_i \in S_i} q(\sigma_i, \hat{s}_{-i})$ since $(s_i, s'_{-i}) \in S_-$. It follows that $q(s_i, \hat{s}_{-i}) = q(s'_i, \hat{s}_{-i})$ for any $s'_i \in S_i$, implying that q is constant with respect to expert i's report. Since this is true for every i, q is a constant mechanism. □

III.6.4 Proof of Proposition III.1

Proof. If in the environment there is no informed expert, then by Lemmas III.3 any DSIC mechanism is constant. Since obviously DM's payoff from the optimal constant mechanism is B, no constant mechanism extracts information. If instead there is an informed expert i, then by Lemma III.4 there exist $q^+ \geq q^-$ such that $q(s) = q_+$ if $s_i \in S_i^R$, and $q(s) = q_-$ if

$s_i \in S_i^L$. Thus

$$\begin{aligned}
-\sum_{s\in S} p(s)e(s)q(s) &= -q_+ \sum_{s:s_i\in S_i^R} p(s)e(s) - q_- \sum_{s:s_i\in S_i^L} p(s)e(s) \\
&= -q_+ \sum_{s:e(s)>0} p(s)e(s) - q_- \sum_{s:e(s)<0} p(s)e(s) \\
&\leq -q_+ \sum_{s_\in S} p(s)e(s) \leq B.
\end{aligned}$$

Again, q does not extract information. □

III.6.5 Proof of Lemma III.5

Proof. Pick any $s \in S_+$. EPIC at s implies $q(s) \geq q(s')$ for any s' that is contiguous to . If further $s' \in S_+$ then similarly $q(s') \geq q(s)$, implying $q(s) = q(s')$. The same is established with an analogous argument if $s, s' \in S_-$. Thus we have Condition 1.

Suppose $s \in S_+$ and $s' \in S_-$ are close. If they are also contiguous then Condition 2 is the immediate consequence of ex post IC at s. If s and s' differ in two entries then there is s'' that is contiguous to s and to s'. EPIC at s implies $q(s) \geq q(s'')$, and EPIC at s' implies $q(s'') \geq q(s')$. It follows that $q(s) \geq q(s')$. Thus we have Condition 2.

Now we show the second part of the lemma. Suppose there exists q satisfying Conditions 1 and 2. Define $\tilde{q}$ such that:

- For any s where $p(s) > 0$: $\tilde{q}(s) = q(s)$.
- For any s where $p(s) = 0$:
  - $\tilde{q}(s) = 1$, if there does not exist $s' \in S_+$ that is contiguous to s.
  - $\tilde{q}(s) = \min\Big\{q(s') : s' \in S_+ \text{ and } s' \text{ is contiguous to } s\Big\}$, if there exists some $s' \in S_+$ that is contiguous to s.

To show that $\tilde{q}$ is EPIC, pick any $s \in S_p$, and s' that is contiguous to s. First suppose $s \in S_+$. If $p(s') > 0$ then Condition 1 or 2 imply $\tilde{q}(s) = q(s) \geq q(s') = \tilde{q}(s')$. If $p(s') = 0$ then by construction $\tilde{q}(s') \leq q(s)$. Thus EPIC holds at s if $s \in S_+$. Now suppose $s \in S_-$. If $p(s') > 0$ then Condition 1 or 2 imply $\tilde{q}(s) = q(s) \leq q(s') = \tilde{q}(s')$. If $p(s') = 0$ and there does not exist $s'' \in S_+$ that is contiguous to s', then by construction $\tilde{q}(s') = 1 \geq \tilde{q}(s)$. If $p(s') = 0$ and there exists some $s'' \in S_+$ that is contiguous to s', then any such s'' is close to s, and hence by Condition 2 we have $\tilde{q}(s) \leq \tilde{q}(s'')$, implying $\tilde{q}(s) \leq \tilde{q}(s') = \min\Big\{q(s'') : s'' \in S_+ \text{ and } s'' \text{ is contiguous to } s'\Big\}$. Thus EPIC holds at s if $s \in S_-$. □

III.6.6 Proof of Proposition III.2

Proof. **"If" direction:** Suppose $v(S_p) \leq 0$ and there is a Type A cluster C. Clearly $B = 0$. Construct q such that $q(s) = 1$ for every $s \in C$ and $q(s) = 0$ for every $s \in S_p - C$. Pick any two contiguous states $s, s' \in S_p$ where $\text{sgn}(e(s)) = \text{sgn}(e(s'))$. Condition C1 implies s and s' are both in C or both in $S_p - C$. In either case $q(s) = q(s')$ by construction of q. Thus Condition 1 in Lemma III.5 is satisfied. Now pick any two close states $s, s' \in S_p$ where $s \in S_-$ and $s' \in S_+$. If they are both in C or both in $S_p - C$ then $q(s) = q(s')$. If $s \in S_p - C$ and $s' \in C$ then $q(s') = 1 > 0 = q(s)$. Condition C2(A) rules out the possibility that $s \in C$ and $s' \in S_p - C$. Thus $q(s') \geq q(s)$, and Condition 2 in Lemma III.5 is satisfied. It follows from Lemma III.5 that there is an EPIC mechanism $\tilde{q}$ where $\tilde{q}(s) = q(s)$ for every $s \in S_p$. DM's expected payoff from $\tilde{q}$ is the same as that from q, which is equal to $v(C) > 0 = B$.

Suppose $v(S_p) \geq 0$ and there is a Type B cluster. In this case $B = v(S_p)$. Construct q such that $q(s) = 0$ for every $s \in C$ and $q(s) = 1$ for every $s \in S_p - C$. Using an analogoius argument as above we can show the existence of an EPIC mechanism $\tilde{q}$ where $\tilde{q}(s) = q(s)$ for any $s \in S_p$. DM's payoff from $\tilde{q}$ is $v(S_p) - v(C) > v(S_p) = B$. This concludes the proof of the "if" direction.

"Only if" direction: Suppose $v(S_p) \leq 0$ and there does not exist a Type A cluster. Let q be the optimal EPIC mechanism for DM.

Define a relation $\sim$ on S_p such that $s \sim s'$ if there is a finite sequence $(s^0, ..., s^n)$ of elements in S_p where (1) $s^0 = s$ and $s^n = s'$, (2) $\text{sgn}(e(s^0)) = ... = \text{sgn}(e(s^n))$, (3) s^k is contiguous to s^{k-1} for every $k = 1, ..., n$. It is straightforward to verify that $\sim$ is an equivalent relation, and hence it partitions S_p into a set $\mathscr{K}$ of equivalence classes. For any $K \in \mathscr{K}$ and $s, s' \in K$, Lemma III.5 implies that $q(s) = q(s')$ because q is EPIC; we use $q(K)$ to denote this probability. Also for any $K \in \mathscr{K}$ and $s, s' \in K$ we have $\text{sgn}(e(s)) = \text{sgn}(e(s'))$; we use $\text{sgn}(K)$ to denote this sign.

For any $K \in \mathscr{K}$ define $R(K) := \{K' \in \mathscr{K} : K' \text{ is close to } K, \text{sgn}(K') \neq \text{sgn}(K)\}$. Pick any $K \in \mathscr{K}$ where $\text{sgn}(K) = +$. Suppose $R(K)$ is empty and $q(K) > 0$. Consider $\tilde{q}$ where $\tilde{q}$ differ from q only in that $\tilde{q}(K) = 0$. It is straightforward to verify that q being EPIC implies $\tilde{q}$ is EPIC. Moreover DM's payoff from $\tilde{q}$ is equal to his payoff from q plus $-q(K)v(K)$, which is strictly higher than his payoff from q because $v(K) < 0$ given $\text{sgn}(K) = +$, contradicting the assumption that q is the optimal EPIC mechanism. Thus $q(K) = 0$ if $R(K)$ is empty.

Now suppose $R(K)$ is nonempty. By Condition 2 in Lemma III.5, $q(K) \geq q(K')$ for every $K' \in R(K)$ because $K \subset S_+$ and $K' \subset S_-$. Suppose $q(K) > \max_{\kappa \in R(K)} q(\kappa)$. Construct $\tilde{q}$ that differs from q only in that $q(K) = \max_{\kappa \in R(K)} q(\kappa)$. Using a similar argument as in the previous case we can show that $\tilde{q}$ is EPIC and gives DM a higher payoff, a contradiction.

Thus $q(K) = \max_{\kappa \in R(K)} q(\kappa)$.

Similarly, for any $K \in \mathscr{K}$ where $\text{sgn}(K) = -$ we can establish that $q(K) = 1$ if $R(K)$ is empty, or $q(K) = \min_{\kappa \in R(K)} q(\kappa)$ otherwise.

Define a relation $\approx$ on $\mathscr{K}$ such that $K \approx K'$ if there is a finite sequence $(K^0, ..., K^n)$ of elements in $\mathscr{K}$ where (1) $K^0 = K$ and $K^n = K'$, (2) $K^k \in R(K^{k-1})$ for $k = 1, ..., n$. It is straightforward to verify that $\approx$ is an equivalent relation, and hence it partitions $\mathscr{K}$ into a set $\mathscr{B}$ of equivalence classes.

Consider any $B \in \mathscr{B}$ where $B = \{K\}$ for some $K \in \mathscr{K}$. Observe that $\text{sgn}(K) = +$, for otherwise K would be a Type A cluster. It then immediately follows that $q(K) = 0$.

Now consider any $B \in \mathscr{B}$ where B is not a singleton. It follows that $\{K \in B : \text{sgn}(K) = -\}$ is nonempty. Fix some $K^* \in \{K \in B : \text{sgn}(K) = -\}$ where $q(K^*) = \max_{K \in B: \text{sgn}(K) = -} q(K)$. Denote $N^0 = \{K^*\}$. For any $i \geq 1$, define $N^i := \Big\{K \in B - \bigcup_{j=0}^{i-1} N^j : K \in R(K') \text{ for some } K' \in N^{i-1}\Big\}$ if i is odd, or $N^i = \Big\{K \in B - \bigcup_{j=0}^{i-1} N^j : K \in R(K') \text{ for some } K' \in N^{i-1} \text{ and } q(K) = q(K^*)\Big\}$ if i is even. It is straightforward to verify that for any $K \in N^i$, $\text{sgn}(K) = -$ if i is even or $\text{sgn}(K) = +$ if i is odd, and that $N^i \subset B$ for any i. Let k be the highest number where N^k is nonempty. Observe that by definition $q(K) = q(K^*)$ for any $K \in N^i$ where i is even. Fix $i \leq k$ where i is odd, and pick any $K \in N^i$. By definition $K \in R(K')$ for some $K' \in N^{i-1}$. Since i is even, we have $\text{sgn}(K) = +$ and $\text{sgn}(K') = -$. Thus $q(K) \geq = q(K') = q(K^*)$. Since $R(K) \subset B$, we have $q(K) = \max_{\kappa \in R(K)} q(\kappa) \leq \max_{\kappa \in B, \text{sgn}(\kappa) = -} q(\kappa) = q(K^*)$. Therefore $q(K) = q(K^*)$. We have thus established that $q(K) = q(K^*)$ for every $K \in N^0 \cup ... \cup N^k$.

Define $M := \Big\{s : s \in K \text{ for some } K \in N^0 \cup ... \cup N^k\Big\}$. Observe that by construction there is no $s \in M \cap S_-$ that is close to some $s' \in (S_p - M) \cap S_+$. Therefore $v(M) \leq 0$, for otherwise M would be a Type A cluster. Define $L := \Big\{K \in B - \bigcup_{j=0}^{k} N^k : K \in R(K') \text{ for some } K' \in N^k\Big\}$. If S is empty then $v(M) \leq 0$ implies there is an optimal EPIC mechanism $\tilde{q}$ where $\tilde{q}(s) = 0$ if $s \in M$ or $\tilde{q}(s) = q(s)$ if $s \in S_p - M$. If S is nonempty then $\text{sgn}(K) = -$ and $q(K) < q(K^*)$ for every $K \in L$. Consider mechanism $\hat{q}$ where $\hat{q}(s) = \max_{K \in L} q(K)$ if $s \in M$ or $\hat{q}(s) = q(s)$ otherwise. Clearly $\hat{q}(s)$ is also EPIC, and moreover DM's payoff from $\hat{q}(s)$ is equal to his payoff from $q(s)$ plus $\Big(q(K^*) - \max_{K \in L} q(K)\Big)v(M)$. That q is optimal implies $v(M) = 0$. Thus $\hat{q}$ is also an optimal EPIC mechanism. Moreover note that $\hat{q}(K^*) = \max_{K \in B: \text{sgn}(K) = -} \hat{q}(K)$. Applying the same argument used in the previous paragraph and this paragraph on $\hat{q}$ still starting with K^*, we can show that there is a subset of B that is strictly larger than $N^0 \cup ... \cup N^k$ (because there is now some $K' \in L$ such that $\tilde{q}(K') = \tilde{q}(K^*)$) such that $q(K) = q(K^*)$ for every K in that subset. Repeatedly applying the argument and we will eventually reach some optimal EPIC $\overline{q}$ where $\overline{q}(K) = \overline{q}(K^*)$ for

every $K \in B$. Define $\overline{M} := \{s : s \in K \text{ for some } K \in B\}$. Observe that $v(\overline{M}) \leq 0$ for otherwise $\overline{M}$ would be a Type A cluster. Thus there is some optimal EPIC $\tilde{q}$ where $\tilde{q}(s) = 0$ if $s \in \overline{M}$ or $\tilde{q}(s) = q(s)$ if $s \in S_p - M$. Applying the same argument to every $B \in \mathscr{B}$, we have established the existence of an optimal EPIC mechanism that is equal to 0 on S_p. Thus DM cannot beat the benchmark. The proof for the case where $v(S_p) \geq 0$ and there does not exist a Type B cluster is similar.

□

III.6.7 Proof of Proposition III.3

Proof. **"If"**. Suppose the environment is not rich. If $N = 2$ then non-richness implies in all states that have positive probability the optimal option is always the same. Thus always choosing that optimal option is first-best and is trivially thruthfully implementable.

Consider $N > 3$. Take any choice rule q^* such that:

- For positive-probability s: $q^*(s) = 1$ if $e(s) < 0$ or $q^*(s) = 0$ if $e(s) > 0$.
- For zero-probability s: If s differs from some positive-probability s' only in one signal then $q^*(s) = q^*(s')$.

Non-richness implies that if positive-probability states $\hat{s}$ and $\tilde{s}$ are in the one-signal-difference neighborhood of a zero-probability state s then $\text{sgn}(e(\hat{s})) = \text{sgn}(e(\tilde{s}))$ and hence $q^*(\hat{s}) = q^*(\tilde{s})$. Thus q^* is well-defined. q^* is also first best by construction. Clearly q^* is truthfully implementable because any expert i is indifferent between all messages regardless of his signal.

"Only if": We show the contrapositive. Pick any first-best rule q^*. Suppose the environment is rich. There are s and s' such that $p(s) > 0$, $p(s') > 0$, $e(s) > 0$, $e(s') < 0$, and s is different from s' only in one or two signals. It follows that $q^*(s) = 0$ and $q^*(s') = 1$. First consider the case that s and s' differ only in the ith signal. Suppose all players are truthful except i who deviates from being truthful to always reporting s_i regardless of his signal. It follows that in s' the inferior option[8] B is chosen, which in turn implies the deviation strictly decreases DM's payoff (since $p(s') > 0$), or equivalently strictly increases expert i's payoff. Thus q^* is not truthfully implementable.

Now suppose s and s' differ only in the signals of experts i and j. Let s_{-ij} denote the signals of experts other than i and j under s. Consider $Q := q^*(s'_i, s_j, s_{-ij})$. Suppose $Q > 0$, then if expert i unilaterally deviates from being truthful to always reporting s'_i then in state s the inferior option A is chosen with positive probability, implying a strict decrease in DM's

[8]Inferior by DM's preference.

payoff from its first-best level and hence a strict increase in expert i's payoff. Similarly, if $Q < 1$ then expert j unilaterally deviating to always reporting s_j has the same effect since in state s' the inferior option B is chosen with positive probability. Thus q^* is not truthfully implementable. □

III.6.8 Proof of Lemma III.6

Proof. **"If" direction:** Fix expert i. Suppose the other experts use the correlated strategy of jointly reporting s_{-i} with probability $z(s_{-i}|\hat{s}_{-i})$ conditional on true signal profile $\hat{s}_{-i}$. Given signal s_i, expert i's expected payoff from truth-telling is equal to

$$\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i, \hat{s}_{-i}) \sum_{s_{-i}\in S_{-i}} z(s_{-i}|\hat{s}_{-i}) q(s_i, s_{-i})$$

whereas his payoff from misreporting as having signal s_i' is

$$\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i, \hat{s}_{-i}) \sum_{s_{-i}\in S_{-i}} z(s_{-i}|\hat{s}_{-i}) q(s_i', s_{-i}).$$

The change in payoff from deviating from truth-telling to misreporting is thus

$$D = -\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i, \hat{s}_{-i}) \sum_{s_{-i}\in S_{-i}} z(s_{-i}|\hat{s}_{-i}) \Big(q(s_i, s_{-i}) - q(s_i', s_{-i}) \Big). \tag{III.1}$$

Choose $\underline{s}_{-i} \in \operatorname{argmin}_{s_{-i}\in S_{-i}} \Big(q(s_i, s_{-i}) - q(s_i', s_{-i}) \Big)$ and $\overline{s}_{-i} \in \operatorname{argmax}_{s_{-i}\in S_{-i}} \Big(q(s_i, s_{-i}) - q(s_i', s_{-i}) \Big)$. Observe D is the highest when $z(\overline{s}_{-i}|\hat{s}_{-i}) = 1$ for any $\hat{s}_{-i}$ where $h(s_i, \hat{s}_{-i}) < 0$ and $z(\underline{s}_{-i}|\hat{s}_{-i}) = 1$ for any $\hat{s}_{-i}$ where $h(s_i, \hat{s}_{-i}) > 0$. Thus

$$\begin{aligned}
D \leq & -\Big(q(s_i, \overline{s}_{-i}) - q(s_i', \overline{s}_{-i}) \Big) \sum_{\hat{s}_{-i}\in S_{-i}}^{h(s_i,\hat{s}_{-i})<0} h(s_i, \hat{s}_{-i}) \\
& - \Big(q(s_i, \underline{s}_{-i}) - q(s_i', \underline{s}_{-i}) \Big) \sum_{\hat{s}_{-i}\in S_{-i}}^{h(s_i,\hat{s}_{-i})\geq 0} h(s_i, \hat{s}_{-i}) \\
= & -\Big(q(s_i, \overline{s}_{-i}) - q(s_i', \overline{s}_{-i}) \Big) \underline{\alpha}_i(s_i) - \Big(q(s_i, \underline{s}_{-i}) - q(s_i', \underline{s}_{-i}) \Big) \overline{\alpha}_i(s_i) \\
\leq & \, 0
\end{aligned}$$

where the last inequality is by assumption. It follows that truth-telling is best responding. Since i, s_i, s_i' and z are arbitrarily chosen, we conclude that q is interim dominant strategy

incentive compatible.

"Only if" direction: Suppose q is interim dominant strategy incentive compatible. Pick any expert i, $s_i, s_i' \in S_i$ and $s_{-i}, s'_{-i} \in S_j$. Suppose experts other than i use the (correlated) strategy profile of reporting s_{-i} with probability $z(\hat{s}_{-i})$ and s'_{-i} with probability $1 - z(\hat{s}_{-i})$ conditional on joint signal profile $\hat{s}_{-i}$.

Since q is strategy proof, reporting s_i instead of s_i' given signal s_i is a best response for i. Thus

$$\begin{aligned}&\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i,\hat{s}_{-i})\Big(z(\hat{s}_{-i})q(s_i,s_{-i})+(1-z(\hat{z}_{-i}))q(s_i,s'_{-i})\Big)\\ \geq &\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i,\hat{s}_{-i})\Big(z(\hat{s}_{-i})q(s_i',s_{-i})+(1-z(\hat{z}_{-i}))q(s_i',s'_{-i})\Big)\end{aligned}$$

Denote $y := q(s_i, s_{-i}) - q(s_i', s_{-i})$ and $y' := q(s_i, s'_{-i}) - q(s_i', s'_{-i})$. The above inequality can be rearranged as:

$$y \sum_{\hat{s}_{-i}\in S_{-i}} h(s_i,\hat{s}_{-i})z(\hat{s}_{-i}) + y' \sum_{\hat{s}_{-i}\in S_{-i}} h(s_i,\hat{s}_{-i})(1-z(\hat{s}_{-i})) \geq 0. \tag{III.2}$$

interim dominant strategy incentive compatibility requires that this inequality holds for any $(z(\hat{s}_{-i}))_{s_{-i}\in S_{-i}} \in [0,1]^{|S_{-i}|}$. Observe that $\sum_{\hat{s}_{-i}\in S_{-i}} h(s_i,\hat{s}_{-i})z(\hat{s}_{-i})$ can take any value that is between $\underline{\alpha}_i(s_i)$ and $\overline{\alpha}_i(s_i)$ given the appropriate choice of z. It follows that inequality III.2 holds if and only if

$$yk + y'(\beta_i(s_i) - k) \geq 0 \qquad \forall k \in [\underline{\alpha}_i(s_i), \overline{\alpha}_i(s_i)]. \tag{III.3}$$

An immediate implication is that $\underline{\alpha}_i(s_i)y + \overline{\alpha}_i(s_i)y' \geq 0$ by setting $k = \underline{\alpha}_i(s_i)$. This concludes the proof. □

III.6.9 Proof of Proposition III.4

We first state and prove a lemma that is helpful for the proof of the proposition. For expert i, let S_i^+ denote the set of $s_i \in S_i$ where $\beta_i(s_i) \geq 0$ and S_i^- the analogous set where $\beta_i(s_i) < 0$.

Let Σ denote $\{+,-\}^n$. Impose a (partial) ordering $\geq$ on Σ such that $\sigma \geq \sigma'$ if $\sigma_i' = -$ for any i where $\sigma_i = -$. For $\sigma \in \Sigma$ let $S(\sigma)$ denote $\times_i S_i^{\sigma_i}$. Note that $\{S(\sigma)\Big|\sigma \in \Sigma\}$ is a partition of S.

Lemma III.11. *For any interim dominant strategy incentive compatible mechanism q there is another interim dominant strategy incentive compatible mechanism r which gives the ex-*

perts the same payoff and moreover:

1. $r(s) = r(s')$ *if* $s, s' \in S(\sigma)$ *for some* $\sigma \in \Sigma$.

2. $r(s) \geq r(s')$ *if* $s \in S(\sigma)$ *and* $s' \in S(\sigma')$ *where* $\sigma \geq \sigma'$.

Proof. Pick any i, $s_i \in S_i$ where $\beta_i(s_i) > 0$. For any $s'_i \in S_i$ and $s_{-i} \in S_{-i}$, Lemma III.6 implies (by setting $s_{-i} = s'_{-i}$) that $\beta_i(s_i)\Big(q(s_i, s_{-i}) - q(s'_i, s_{-i})\Big) \geq 0$, which in turn implies $q(s_i, s_{-i}) \geq q(s'_i, s_{-i})$. If in addition $\beta_i(s'_i) > 0$ then by symmetry the reverse inequality also holds, implying $q(s_i, s_{-i}) = q(s'_i, s_{-i})$. Similarly if $\beta_i(s_i) < 0$ and $\beta_i(s'_i) < 0$ then $q(s_i, s_{-i}) = q(s'_i, s_{-i})$ for any s_{-i}.

For each i let $S_i^{>}/S_i^{=}/S_i^{<}$ be the set of $s_i \in S_i$ where $\beta_i(s_i) >=< 0$. Note that $S_i^{<} = S_i^{-}$ and $S_i^{+} = S_i^{>} \cup S_i^{=}$. We have shown in the previous paragraph that for any s_{-i},

a. If $s_i \in S_i^{>}$ then $q(s_i, s_{-i})$ is a constant which we denote as $\overline{q}_i(s_{-i})$.

b. If $s_i \in S_i^{<}$ then $q(s_i, s_{-i})$ is a constant which we denote as $\underline{q}_i(s_{-i})$.

c. $\overline{q}_i(s_{-i}) \geq \underline{q}_i(s_{-i})$

Now we construct a sequence of mechanisms, starting with $r^0 = q$, such that for any $i \leq n$, r^i is obtained from r^{i-1} by the following protocol:

- If $s_i \notin S_i^{=}$: $r^i(s) = r^{i-1}(s)$.
- If $s_i \in S_i^{=}$ and $S_i^{>} \neq \emptyset$: $r^i(s) = r^{i-1}(s'_i, s_{-i})$ where s'_i is chosen from $S_i^{>}$ independent of s_i and s_{-i}.
- If $s_i \in S_i^{=}$, $S_i^{>} = \emptyset$ and $S_i^{<} \neq \emptyset$: $r^i(s) = r^{i-1}(s'_i, s_{-i})$ where s'_i is chosen from $S_i^{<}$ independent of s_i and s_{-i}.
- If $s_i \in S_i^{=}$, $S_i^{>} = \emptyset$ and $S_i^{<} = \emptyset$: $r^i(s) = r^{i-1}(s'_i, s_{-i})$ for some s'_i chosen independent of s_i and s_{-i}.

Stop at $r^n = r$.

Consider the inductive hypotheses for S:

(H1) For every $i \leq l$ and s_{-i},

(a) $r^l(\cdot, s_{-i})$ is constant over $S_i^{>} \cup S_i^{=}$.

(b) $r^l(\cdot, s_{-i})$ is constant over $S_i^{<}$.

(c) $r^l(s_i, s_{-i}) \geq r^l(s'_i, s_{-i})$ if $s_i \in S_i^{>} \cup S_i^{=}$ and $s'_i \in S_i^{<}$.

(H2) For every $i > l$, and s_{-i},

(a) $r^l(\cdot, s_{-i})$ is constant over $S_i^>$.

(b) $r^l(\cdot, s_{-i})$ is constant over $S_i^<$.

(c) $r^l(s_i, s_{-i}) \geq r^l(s_i', s_{-i})$ if $s_i \in S_i^>$ and $s_i' \in S_i^<$.

(H3) The experts' payoff from r^l is the same to that from q.

(H4) r^l is interim dominant strategy incentive compatible.

Consider $l = 0$. H3 and H4 are true by assumption. H1 is vacuously true. For any $i = 1, ..., n$ and s_{-i}, $r^0(s_i, s_{-i}) = \overline{q}_i(s_{-i})$ if $s_i \in S_i^>$, or $r^0(s_i, s_{-i}) = \underline{q}_i(s_{-i})$ if $s_i \in S_i^<$. Moreover $\overline{q}_i(s_{-i}) \geq \underline{q}_i(s_{-i})$. Thus H2 is true for $j = 0$.

Suppose H1-H4 are true for some $l = j < n$. We first show that H1 holds for $l = j + 1$. Fix any $i < j+1$ and s_{-i}. Let s_{-i}^k denote the signal of expert $k \neq i$ under s_{-i}. If $s_{-i}^{j+1} \notin S_{j+1}^=$ then $r^{j+1}(\cdot, s_{-i}) = r^j(\cdot, s_{-i})$, which immediately implies H1 for $l = j + 1$ in this case, given that H1 holds for $l = j$. If $s_{-i}^{j+1} \in S_{j+1}^=$ and $S_{j+1}^> \neq \emptyset$, then for any s_i, $r^{j+1}(s_i, s_{-i}) = r^j(s_i, s_{j+1}', s_{-(i,j+1)})$ where s_{j+1}' is chosen from $S_{j+1}^>$. Since by H1 for $l = j$, $r^j(\cdot, s_{j+1}', s_{-(i,j+1)})$ is constant over $S_i^> \cup S_i^=$ or $S_i^<$, and moreover $r^j(s_i, s_{j+1}', s_{-(i,j+1)}) \geq r^j(s_i', s_{j+1}', s_{-(i,j+1)})$ if $s_i \in S_i^> \cup S_i^=$ and $s_i' \in S_i^<$, H1 for $l = j + 1$ follows in this case. Similarly H1 for $l = j + 1$ follows in the case $s_{-i}^{j+1} \in S_{j+1}^=$, $S_{j+1}^> = \emptyset$ and $S_{j+1}^< \neq \emptyset$. If $s_{-i}^{j+1} \in S_{j+1}^=$ and $S_{j+1}^>, S_{j+1}^<$ are both empty, then $r^{j+1}(s_i, s_{-i}) = r^j(s_i, s_{j+1}', s_{-(i,j+1)})$ for some s_{j+1}'. Note that, by H1 for $l = j$, for any s_{j+1}', $r^j(\cdot, s_{j+1}', s_{-(i,j+1)})$ is constant over $S_i^> \cup S_i^=$ or $S_i^<$, and moreover $r^j(s_i, s_{j+1}', s_{-(i,j+1)}) \geq r^j(s_i', s_{j+1}', s_{-(i,j+1)})$ if $s_i \in S_i^> \cup S_i^=$ and $s_i' \in S_i^<$. This consequently implies H1 for $l = j + 1$ in this case.

Now consider $i = j + 1$. Fix any s_{-j+1}. Since by H2 for $l = j$, $r^j(\cdot, s_{-j})$ is constant over $S_{j+1}^>$ or $S_{j+1}^<$ and moreover $r^j(s_i, s_{-i}) \geq r^j(s_i', s_{-i})$ if $s_i \in S_i^>$ and $s_i' \in S_i^<$. the construction of r^{j+1} from r^j immediately implies that $r^j(\cdot, s_{-i})$ is constant over $S_i^> \cup S_i^=$ or $S_i^<$, and moreover $r^{j+1}(s_i, s_{-i}) \geq r^{j+1}(s_i', s_{-i})$ if $s_i \in S_i^> \cup S_i^=$ and $s_i' \in S_i^<$. We have thus fully established H1 for $l = j + 1$.

H2 for $l = j + 1$ is is established using a similar argument to how H1 for $l = j + 1$ is established in cases where $i < j+1$, except that $S_i^> \cup S_i^=$ is correspondingly replaced by $S_i^>$.

To show that H3 holds for $l = j + 1$ it is sufficient to show that the experts' common payoff π^{j+1} from r^{j+1} is the same from their common payoff π^j from r^j. Observe that

$$\begin{aligned} &\pi^{j+1} - \pi^j \\ &= \sum_{s_{j+1} \in S_{j+1}^=} \sum_{s_{-(j+1)} \in S_{-(j+1)}} h(s_{j+1}, s_{-(j+1)}) \Big(r^{j+1}(s_{j+1}, s_{-(j+1)}) - r^j(s_{j+1}, s_{-(j+1)}) \Big). \qquad (*) \end{aligned}$$

Since H4 holds for $l = j$, Lemma III.6 implies that for any $s_{j+1} \in S^{=}_{j+1}$, $s'_{j+1} \in S_{j+1}$, and $s_{-(j+1)}, s'_{-(j+1)} \in S_{-(j+1)}$,

$$\begin{aligned}&\underline{\alpha}_{j+1}(s_{j+1})\Big(r^j(s_{j+1}, s_{-(j+1)}) - r^j(s'_{j+1}, s_{-(j+1)})\Big)\\ &+ \overline{\alpha}_{j+1}(s_{j+1})\Big(r^j(s_{j+1}, s'_{-(j+1)}) - r^j(s'_{j+1}, s'_{-(j+1)})\Big) \geq 0.\end{aligned}$$

Since $\overline{\alpha}_{j+1}(s_{j+1}) = -\underline{\alpha}_{j+1}(s_{j+1}) > 0$, the above inequality simplifies to

$$r^j(s_{j+1}, s_{-(j+1)}) - r^j(s'_{j+1}, s_{-(j+1)}) \leq r^j(s_{j+1}, s'_{-(j+1)}) - r^j(s'_{j+1}, s'_{-(j+1)}).$$

By symmetry between $s_{-(j+1)}$ and $s'_{-(j+1)}$ we also have the reverse inequality. Therefore we have

$$r^j(s'_{j+1}, s_{-(j+1)}) - r^j(s_{j+1}, s_{-(j+1)}) = r^j(s'_{j+1}, s'_{-(j+1)}) - r^j(s_{j+1}, s'_{-(j+1)}).$$

It follows that $r^j(s'_{j+1}, s_{-(j+1)}) - r^j(s_{j+1}, s_{-(j+1)})$ is independent of $s_{-(j+1)}$; we can thus denote the difference as $\delta(s'_{j+1}, s_{j+1})$.

Observe that for any $s_{j+1} \in S^{=}_{j+1}$, by construction $r^{j+1}(s_{j+1}, s_{-(j+1)}) = r^j(s'_{j+1}, s_{-(j+1)})$ for some s'_{j+1} that is independent of s_{j+1} and $s_{-(j+1)}$. Therefore equation (*) becomes

$$\begin{aligned}&\pi^{j+1} - \pi^j\\ =&\sum_{s_{j+1} \in S^{=}_{j+1}} \delta(s'_{j+1}, s_{j+1}) \sum_{s_{-(j+1)} \in S_{-(j+1)}} h(s_{j+1}, s_{-(j+1)})\\ =&\sum_{s_{j+1} \in S^{=}_{j+1}} \delta(s'_{j+1}, s_{j+1}) \times 0\\ =&0.\end{aligned}$$

Thus H3 holds for $l = j + 1$.

Now we show that H4 holds for $l = j+1$. Pick any expert i, $s_i, s'_i \in S_i$ and $s_{-i}, s'_{-i} \in S_{-i}$. By Lemma III.6 it is sufficient to verify:

$$\underline{\alpha}_i(s_i)\Big(r^{j+1}(s_i, s_{-i}) - r^{j+1}(s'_i, s_{-i})\Big) + \overline{\alpha}_i(s_i)\Big(r^{j+1}(s_i, s'_{-i}) - r^{j+1}(s'_i, s'_{-i})\Big) \geq 0. \qquad \text{(III.4)}$$

If $i = j + 1$ and $s_i, s'_i \notin S^{=}_{j+1}$, then $r^{j+1}(\hat{s}) = r^j(\hat{s})$ for any $\hat{s} \in \{s_i, s'_i\} \times \{s_{-i}, s'_{-i}\}$. Thus inequality III.4 holds because the analogous inequality holds for these states under r^j due to H4 holding for $l = j$. Now suppose $i = j + 1$ and $s_i \in S^{=}_{j+1}$. If $s'_i \in S^{>}_{j+1} \cup S^{=}_{j+1}$ then $r^{j+1}(s_i, \hat{s}_{-i}) = r^{j+1}(s'_i, \hat{s}_{-i})$ for any $\hat{s}_{-i} \in S_{-i}$ by construction and H1 or H2, implying that the

LHS of inequality III.4 is 0. Thus the inequality holds. If $s_i' \in S_{j+1}^{<}$ and $S_{j+1}^{>}$ is empty then again $r^{j+1}(s_i, \hat{s}_{-i}) = r^{j+1}(s_i', \hat{s}_{-i})$ for any $\hat{s}_{-i} \in S_{-i}$ by construction and H1 or H2, implying that inequality III.4 holds. If $s_i' \in S_{j+1}^{<}$ and $S_{j+1}^{>}$ is nonempty, then $r^{j+1}(s_i, s_{-i}) = r^j(s_i'', s_{-i})$ and $r^{j+1}(s_i, s_{-i}') = r^j(s_i'', s_{-i}')$ for some $s_i'' \in S_{j+1}^{>}$. Recall that we have shown (in the proof of H3) that $r^j(s_i, \cdot) - r^j(\hat{s}_i, \cdot)$ for any given $\hat{s}_i \in S_{j+1}$ is constant over $S_{-(j+1)}$. Thus we have

$$\begin{aligned}
r^{j+1}(s_i, s_{-i}) - r^{j+1}(s_i', s_{-i}) &= r^j(s_i'', s_{-i}) - r^j(s_i', s_{-i}) \\
&= \Big(r^j(s_i'', s_{-i}) - r^j(s_i, s_{-i})\Big) + \Big(r^j(s_i, s_{-i}) - r^j(s_i', s_{-i})\Big) \\
&= \Big(r^j(s_i'', s_{-i}') - r^j(s_i, s_{-i}')\Big) + \Big(r^j(s_i, s_{-i}') - r^j(s_i', s_{-i}')\Big) \\
&= r^j(s_i'', s_{-i}') - r^j(s_i', s_{-i}') \\
&= r^{j+1}(s_i, s_{-i}') - r^{j+1}(s_i', s_{-i}').
\end{aligned}$$

It follows that, also substituting in $\overline{\alpha}_i(s_i) = -\underline{\alpha}_i(s_i) > 0$, that the LHS of inequality III.4 is equal to 0. Thus the inequality always holds if $i = j + 1$.

Suppose $i \neq j + 1$. It follows from the construction that $r^{j+1}(s_i, s_{-i}) = r^j(s_i, \hat{s}_{-i})$ and $r^{j+1}(s_i', s_{-i}) = r^j(s_i, \hat{s}_{-i})$ for some $\hat{s}_{-i}$. Similarly $r^{j+1}(s_i, s_{-i}') = r^j(s_i, \tilde{s}_{-i})$ and $r^{j+1}(s_i', s_{-i}') = r^j(s_i, \tilde{s}_{-i})$ for some $\tilde{s}_{-i}$. Thus we have

$$\begin{aligned}
&\underline{\alpha}_i(s_i)\Big(r^{j+1}(s_i, s_{-i}) - r^{j+1}(s_i', s_{-i})\Big) + \overline{\alpha}_i(s_i)\Big(r^{j+1}(s_i, s_{-i}') - r^{j+1}(s_i', s_{-i}')\Big) \\
=&\underline{\alpha}_i(s_i)\Big(r^j(s_i, \hat{s}_{-i}) - r^j(s_i', \hat{s}_{-i})\Big) + \overline{\alpha}_i(s_i)\Big(r^j(s_i, \tilde{s}_{-i}) - r^j(s_i', \tilde{s}_{-i})\Big) \geq 0
\end{aligned}$$

where the inequality is due to r^j being interim dominant strategy incentive compatible by the inductive hypothesis H4 for $l = j$.

We have thus proved by induction that H1-H4 hold for $l = n$. Therefore $r = r^n$ is interim dominant strategy incentive compatible and expert-payoff equivalent to q.

Given H1, for every i and s_{-i}, $r(\cdot, s_{-i}) = r^n(\cdot, s_{-i})$ is constant over $S_i^+ = S_i^> \cup S_i^=$ or $S_i^- = S_i^<$, and moreover $r(s_i, s_{-i}) \geq r(s_i', s_{-i})$ if $s_i \in S_i^+$ and $s_i' \in S_i^-$. Pick any $s, s' \in S$. There is a sequence of states $s = s^0, s^1, ..., s^n = s'$ where s^i and s^{i-1} can only possibly differ in that $s_i^i = s_i'$ and $s_i^{i-1} = s_i$. If there is some $\sigma \in \Sigma$ where $s, s' \in S(\sigma)$ then for any $i = 1, ..., n$, s_i and s_i' are both in S_i^+ or S_i^-, implying $r(s^{i-1}) = r(s_i, s_{-i}^i) = r(s_i', s_{-i}^i) = r(s^i)$. It follows that $r(s) = r(s')$. If $s \in \sigma$ and $s' \in \sigma'$ where $\sigma \geq \sigma'$ then for any $i = 1, ..., n$, either s_i and s_i' are both in S_i^+ or S_i^-, or $s_i \in S_i^+$ and $s_i' \in S_i^-$, implying $r(s^{i-1}) = r(s_i, s_{-i}^i) \geq r(s_i', s_{-i}^i) = r(s^i)$, which in turn implies $r(s) \geq r(s')$. We have thus established conditions 1 and 2 from the statement of the lemma. □

Now we are read to Prove the Proposition.

Proof. Let q be an interim dominant strategy incentive compatible mechanism and r the mechanism that is payoff-equivalent to q as specified in Lemma III.11. It is sufficient to show that the experts achieve at least the benchmark payoff from r. Since r is constant on each $\sigma \in \Sigma$ we can denote (by abusing notation when no confusion arises) that constant as $r(\sigma)$. Also define $H(\sigma) := \sum_{s \in S(\sigma)} h(s)$.

First suppose S_i^+ and S_i^- are nonempty for every i. For any $i = 1, ..., n$ and $\sigma \in \Sigma$ let $\mu^i(\sigma)$ denote the resulting vector by changing the ith entry of σ to $+$. (If $\sigma_i = +$ already then $\mu^i(\sigma) = \sigma$). For each $i = 0, ..., n$ define functions $\lambda^i : \Sigma \to \Sigma$ such that:

1. $\lambda^0(\sigma) = (-, ..., -)$ for every $\sigma \in \Sigma$.

2. For $i > 0$,

   (a) $\lambda^i(\sigma) = \lambda^{i-1}(\sigma)$ if $\sigma_i = -$.

   (b) $\lambda^i(\sigma) = \mu^i(\lambda^{i-1}(\sigma))$ if $\sigma_i = +$.

For each $i = 0, ..., n$ define mechanism r^i such that $r^i(s) = r(\lambda^i(\sigma))$ where $s \in S(\sigma)$.

Note r^0 is a constant mechanism. Clearly the experts get a weakly higher payoff from this mechanism than the benchmark, because the benchmark is obviously the lowest payoff that the experts can get in *any* constant mechanism. Therefore, to prove the proposition it is sufficient to show that: (1) $r^n = r$, and (2) The experts get a weakly higher payoff in r^i than in r^{i-1} conditional on truth-telling.

To establish $r^n = r$, consider the following inductive hypothesis:

L1 For any σ, the first j entries of $\lambda^j(\sigma)$ and σ agree; the other entries of $\lambda^j(\sigma)$ are "$-$".

L1 is clearly true for $j = 0$. Suppose it is true for $j = i - 1$ for some $i \leq n$. Note that $\lambda^i(\sigma)$ and $\lambda^{i-1}(\sigma)$ may only differ at the ith entry. Moreover, L1 for $j = i - 1$ implies the ith entry of $\lambda^{i-1}(\sigma)$ is "$-$", thus by constructions the ith entry of $\lambda^i(\sigma)$ is "$+$" if and only if $\sigma_i = +$, that is, the ith entries of $\lambda^i(\sigma)$ and σ agree. This observation combined with L1 for $i = j - 1$ implies L1 for $i = j$. Therefor, L1 is true for $j = n$ by induction, which implies that $\lambda^n(\sigma) = \sigma$. It immediately follows that $r^n = r$.

Let π^i denote the experts' common payoff from mechanism r^i conditional on truth-telling. We have

$$\pi^i - \pi^{i-1} = \sum_{\sigma : \sigma_i = +} H(\sigma)\Big(r(\lambda^i(\sigma)) - r(\lambda^{i-1}(\sigma))\Big). \tag{III.5}$$

To show that expression III.5 is nonnegative, we go back temporarily to the interim dominant strategy incentive compatible mechanism r. Suppose experts other than i do the following: Expert $j < i$ report the signal; expert $j > i$ always reports some $s_j \in S_j^-$. It is

easy to verify, using L1, that if s obtains where $s \in S(\sigma)$ for some $\sigma \in \Sigma$ and $s_i \in S_i^+$, then if i reports the true signal s_i the reported state is in $\lambda^i(\sigma)$, whereas if i deviates to reporting some $s_i' \in S_i^-$ the reported state is in $\lambda^{i-1}(\sigma)$. Given any $\sigma_{-i} \in \Sigma_{-i}$ let $S_{-i}(\sigma_{-i})$ denote $\times_{j \neq i} S_j^{\sigma_{-i}(j)}$ where $\sigma_{-i}(j)$ denotes the entry correspond to expert j. Since deviating is not given r is strategy proof, for any $s_i \in S_i^+$ we have

$$\sum_{\sigma:\sigma_i=+} \sum_{s_{-i} \in S_{-i}(\sigma_{-i})} h(s_i, s_{-i}) \Big(r(\lambda^i(\sigma)) - r(\lambda^{i-1}(\sigma)) \Big) \geq 0.$$

It follows that

$$\begin{aligned} 0 &\leq \sum_{\sigma:\sigma_i=+} \sum_{s_i \in S_i^+} \sum_{s_{-i} \in S_{-i}(\sigma_{-i})} h(s_i, s_{-i}) \Big(r(\lambda^i(\sigma)) - r(\lambda^{i-1}(\sigma)) \Big) \\ &= \sum_{\sigma:\sigma_i=+} H(\sigma) \Big(r(\lambda^i(\sigma)) - r(\lambda^{i-1}(\sigma)) \Big) \end{aligned}$$

where the RHS is exactly expression III.5. Therefor $\pi_i - \pi^{i-1}$ is nonnegative for every $i = 1, ..., n$. This completes the proof for the case that S_i^+ and S_i^- are nonempty for every i.

If S_i^+ or S_i^- is empty for some experts, then we reindex the experts so that these experts are assigned the highest indices $m+1, ..., n$. The proof needs to be modified in two places: First, the first m entries of $\lambda^0(\sigma)$ are still "$-$", but the ith entry where $i > m$ is $\phi \in \{+, -\}$ where S_i^ϕ is not empty. Second, the construction of λ^i stops at λ^m. We can show that $r^m = r$ and each r^i gives the experts a higher payoff than r^{i-1} using essentially the same argument as above. This completes the proof. □

III.6.10 Proof of Proposition III.5

Proof. Suppose for environments (S, e, p) and (S, e', p'), $p(s)e(s) = p'(s)e'(s)$ for every $s \in S$. Let mechanism q be a BIC mechanism that extracts information in (S, e, p), then

$$-\sum_{s \in S} p(s)e(s)q(s) > \max\{0, -\sum_{s \in S} p(s)e(s)\} \tag{III.6}$$

and for any $i = 1, ..., N$ and $s_i, s_i' \in S_i$,

$$\sum_{s_{-i} \in S_{-i}} \frac{p(s_i, s_{-i})}{\sum_{\hat{s}_{-i} \in S_{-i}} p(s_i, \hat{s}_{-i})} e(s_i, s_{-i}) q(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} \frac{p(s_i, s_{-i})}{\sum_{\hat{s}_{-i} \in S_{-i}} p(s_i, \hat{s}_{-i})} e(s_i, s_{-i}) q(s_i', s_{-i}) \tag{III.7}$$

where $\Pr(s_{-i}|s_i)$ is the probability of s_{-i} conditional s_i in (S, e, p). It follows from inequality III.6 that

$$\begin{aligned}-\sum_{s\in S} p'(s)e'(s)q(s) &= -\sum_{s\in S} p(s)e(s)q(s)\\ &> \max\{0, -\sum_{s\in S} p(s)e(s)\} = \max\{0, -\sum_{s\in S} p'(s)e'(s)\}.\end{aligned}$$

Thus q extracts information in (S, e', p').

It follows from inequality III.7 that for any $i = 1, ..., N$ and $s_i, s_i' \in S_i$,

$$\begin{aligned}&\sum_{s_{-i}\in S_{-i}} \frac{p'(s_i, s_{-i})}{\sum_{\hat{s}_i\in S_i} p'(\hat{s}_i, s_{-i})} e'(s_i, s_{-i})q(s_i, s_{-i})\\ =&\frac{\sum_{\hat{s}_i\in S_i} p(\hat{s}_i, s_{-i})}{\sum_{\hat{s}_i\in S_i} p'(\hat{s}_i, s_{-i})} \sum_{s_{-i}\in S_{-i}} \frac{p(s_i, s_{-i})}{\sum_{\hat{s}_{-i}\in S_{-i}} p(s_i, \hat{s}_{-i})} e(s_i, s_{-i})q(s_i, s_{-i})\\ \geq&\frac{\sum_{\hat{s}_i\in S_i} p(\hat{s}_i, s_{-i})}{\sum_{\hat{s}_i\in S_i} p'(\hat{s}_i, s_{-i})} \sum_{s_{-i}\in S_{-i}} \frac{p(s_i, s_{-i})}{\sum_{\hat{s}_{-i}\in S_{-i}} p(s_i, \hat{s}_{-i})} e(s_i, s_{-i})q(s_i', s_{-i})\\ =&\sum_{s_{-i}\in S_{-i}} \frac{p'(s_i, s_{-i})}{\sum_{\hat{s}_i\in S_i} p'(\hat{s}_i, s_{-i})} e'(s_i, s_{-i})q(s_i', s_{-i}).\end{aligned}$$

Thus q is BIC in (S, e', p'). □

III.6.11 Proof of Lemma III.8

The following two results will be useful for the proof of Lemma III.8.

Lemma III.12. *For a symmetric mechanism q, let q_k denote $q(s)$ where $\epsilon(s) = k$, and d_k denote $\binom{N-1}{k}(q_k - q_{k+1})$. q is BIC if and only if*

$$\sum_{k=0}^{N-1} h_k d_k \geq 0, \qquad \sum_{k=0}^{N-1} h_{k+1} d_k \leq 0. \tag{III.8}$$

Proof. In a BIC symmetric mechanism, no experts wishes to over-report, thus

$$\sum_{k=0}^{N-1} \Pr(\epsilon(s_{-i}) = k|s_i = 0)e(k)q_{sym}(k) \geq \sum_{k=0}^{N-1} \Pr(\epsilon(s_{-i}) = k|s_i = 0)e(k)q_{sym}(k+1)$$

Multiplying both sides by $\Pr(s_i = 0)$, we have

$$\sum_{k=0}^{N-1} \binom{N-1}{k} p_k e_k q_{sym}(k) \geq \sum_{k=0}^{N-1} \binom{N-1}{k} p_k e_k q_{sym}(k+1).$$

Rearranging, we have

$$h_0 q_{sym}(0) + \sum_{k=1}^{N-1} \Big[\binom{N-1}{k} h_k - \binom{N-1}{k-1} h_{k-1} \Big] q_{sym}(k) - h_{N-1} q_{sym}(N) \geq 0.$$

Similarly, the absence of profitable under-reporting opportunity implies

$$-h_1 q_{sym}(0) + \sum_{k=1}^{N-1} \Big[\binom{N-1}{k-1} h_k - \binom{N-1}{k} h_{k+1} \Big] q_{sym}(k) + h_N q_{sym}(N) \geq 0.$$

For $k = 0, ..., N$ define

$$v(k) := \binom{N-1}{k} h_k - \binom{N-1}{k-1} h_{k-1}$$

and

$$w(k) = \binom{N-1}{k-1} h_k - \binom{N-1}{k} h_{k+1}$$

where $\binom{0}{-1}$ and $\binom{N-1}{N}$ are set to be 0, then a mechanism is BIC if and only if

$$\sum_{k=0}^{N} v(k) q_{sym}(k) \geq 0 \sum_{k=0}^{N} w(k) q_{sym}(k) \geq 0. \tag{III.9}$$

Plugging in the definitions of q_k and d_k, Conditions (III.9) become Conditions (III.8). □

Lemma III.13. *There exists a BIC decreasing mechanism if and only if there exists a non-negative and non-constant vector $(d_k)_{k=0}^{N-1}$ that satisfies Conditions (III.8).*

Proof. The only if direction is immediate. To show the if direction, let (d_k) be a non-negative and non-constant vector that satisfies Conditions III.8. Let $K := \sum_{k=0}^{N-1} \frac{d_k}{\binom{N-1}{k}}$. Clearly (d_k/K) is also a non-negative and non-constant vector that satisfies Condition III.8.

Consider mechanism q where $q_0 = 1$ and $q_{k+1} = q_k - \Big[1/\binom{N-1}{k}\Big](d_k/K)$. It is straightforward to verify that q is a well-defined decreasing mechanism, that is, (q_k) is non-increasing and non-constant, and that $q_k \in [0, 1]$ for every $k = 0, ..., N$. Observe that $\sum_{k=0}^{N-1} \binom{N-1}{k} h_k (q_k - q_{k+1}) = \sum_{k=0}^{N-1} h_k (d_k/K)$ and $\sum_{k=0}^{N-1} \binom{N-1}{k} h_{k+1}(q_{k+1} - q_k) = \sum_{k=0}^{N-1} h_{k+1}(-d_k/K)$. Thus that (d_k/K) satisfies Condition III.8 implies q is BIC. □

Now we are able to prove Lemma III.8.

Proof. Denote $\eta_k := (h_k, h_{k+1})$. By Lemma III.13, there exists a BIC decreasing mechanism if and only if there is a non-zero conical combination of $\eta_k, k = 0, ..., N-1$ that lies in the fourth quadrant of the two-dimensional Cartesian plane. By assumption none of η_k is in the fourth quadrant, therefore a non-zero conical combination of $\eta_k, k = 0, ..., N-1$ lies in the fourth quadrant if and only if there are $i, j \in \{0, ..., N\}$ where η_i is in the third quadrant, η_j is in the first quadrant, and the convex cone spanned by η_i and η_j is either a line or contains the fourth quadrant, which holds, as can be easily verified, if and only if $\frac{h_{i+1}}{h_i} \geq \frac{h_{j+1}}{h_j}$. □

III.6.12 Proof of Proposition III.6

Proof. It is easy to see that no symmetric increasing BIC mechanism that extracts information. So we want to show that no symmetric decreasing BIC mechanism that extracts information.

Recall $h_k = p_k e_k$ where p_k is the probability where some given state s where $\epsilon(s) = k$ obtains, and e_k is the expected payoff from conviction in that state. Hence,

$$\begin{aligned} h_k &= \Pr(s)\Big(x\frac{\Pr(\text{guilty and } s)}{\Pr(s)} - y\frac{\Pr(\text{innocent and } s)}{\Pr(s)}\Big) \\ &= x\pi\alpha^k(1-\alpha)^{N-k} - y(1-\pi)\beta^k(1-\beta)^{N-k}. \end{aligned}$$

It is straightforward to verify that h_k is increasing in k. Thus for any i, j where $h_i, h_{i+1} < 0, h_j, h_{j+1} > 0$ we have $h_{i+1}/h_i < 1 < h_{j+1}/h_j$. Thus by Lemma III.8 there does not exist a decreasing BIC mechanism for the classical Condorcet Jury model. □

III.6.13 Proof of Proposition III.7

Proof. Consider a general symmetric voting rule where the probability of conviction is $\hat{q}_k$ if there are k guilty reports. Consider the strategy profile where juror i reports guilty with probability a_i given a guilty signal and b_i given an innocent signal. Fix any s, s' that differ only at the ith component where $s'_i = 1$ and $s_i = 0$. Let $Q_{s'}$ and Q_s respectively denote the implied probability of conviction in states s' and s, and $\tilde{p}_k$ the probability that jurors other than i report exactly k guilty signals given s_{-i}. Thus we have

$$Q_{s'} = \sum_{k=0}^{N-1} \tilde{p}_k\Big(a_i\hat{q}_{k+1} + (1-a_i)\hat{q}_k\Big)$$

and similarly

$$Q_s = \sum_{k=0}^{N-1} \tilde{p}_k \Big(b_i \hat{q}_{k+1} + (1 - b_i)\hat{q}_k \Big).$$

Thus

$$Q_{s'} - Q_s = (a_i - b_i) \sum_{k=0}^{N-1} \tilde{p}_k (\hat{q}_{k+1} - \hat{q}_k).$$

If $\hat{q}$ is weakly monotone and $a_i - b_i$ has the same sign for all i then $Q_{s'} - Q_s$ have the same sign for any s, s' that differ only at the ith component where $s'_i = 1$ and $s_i = 0$.

For any s let $X(s)$ denote the set of all s' obtained from flipping an innocent signal in s to a guilty signal, and let $X^{-1}(s)$ denote the set of all s' obtained from flipping a guilty signal in s to an innocent signal. It follows from the previous paragraph that if $\hat{q}$ is weakly monotone and $a_i - b_i$ has the same sign for all i then $Q_{s'} - Q_s$ has the same sign for any s, s' where $s' \in X(s)$. Suppose WLOG $Q_{s'} - Q_s \geq 0$ for any s, s' where $s' \in X(s)$. Thus

$$\frac{1}{|X(s)|} \sum_{s' \in X(s)} Q_{s'} - Q_s \geq 0,$$

where, clearly, $|X(s)| = N - k$. Thus

$$\sum_{s' \in X(s)} Q_{s'} \geq (N - k) Q_s.$$

Also observe

$$\sum_{s:\epsilon(s)=k} \sum_{s' \in X(s)} Q_{s'} = (k+1) \sum_{s':\epsilon(s')=k+1} Q_{s'}.$$

It follows that

$$\begin{aligned}
\frac{1}{\binom{N}{k+1}} \sum_{s':\epsilon(s')=k+1} Q_{s'} &= \frac{1}{(k+1)\binom{N}{k+1}} \sum_{s:\epsilon(s)=k} \sum_{s' \in X(s)} Q_{s'} \\
&\geq \frac{1}{(k+1)\binom{N}{k+1}} \sum_{s:\epsilon(s)=k} (N-k) Q_s \\
&= \frac{1}{\binom{N}{k}} \sum_{s:\epsilon(s)=k} Q_s.
\end{aligned}$$

If (a_i, b_i) constitutes an equilibrium, then (Q_s) is a BIC direct mechanism. Then by Lemma III.7, the symmetric mechanism $q_k = \frac{1}{\binom{N}{k}} \sum_{s:\epsilon(s)=k} Q(s)$ is also BIC. The above inequality shows that if $\hat{q}$ is monotone and $a_i - b_i$ has the same sign for all i then q_k must also be monotone, but we earlier observe that q_k cannot be decreasing. It follows that q_k must be increasing, in which case the equilibrium of the original voting game is not worse than no voting. We have thus shown that as long as the voting rule is monotone (not necessarily increasing) and players follow relatively "similar" strategies ($a_i - b_i$ having the same sign for all i) then no equilibrium can be worse no voting.

Now think about any equilibrium (a_i, b_i) of a voting game with monotone rule $\hat{q}_k$. Fix juror i. Let $\tilde{p}_k(s_{-i})$ denote the probability that the other jurors report exactly k guilty signals in equilibrium given s_{-i}. Conditional on $s_i = 0$, the difference in payoff between reporting 1 and reporting 0 for i is

$$
\begin{aligned}
&\Pi(1|s_i = 0) - \Pi(0|s_i = 0) \\
=&\sum_{s_{-i}} \Pr(s_{-i}|s_i = 0) \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) \hat{q}_{k+1} e(s_i = 0, s_{-i}) - \sum_{s_{-i}} \Pr(s_{-i}|s_i = 0) \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) \hat{q}_k e(s_i = 0, s_{-i}) \\
=&\sum_{s_{-i}} \Pr(s_{-i}|s_i = 0) e(s_i = 0, s_{-i}) \Big[\sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) \hat{q}_{k+1} - \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) \hat{q}_k \Big] \\
=&\sum_{s_{-i}} \Pr(s_{-i}|s_i = 0) e(s_i = 0, s_{-i}) \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) (\hat{q}_{k+1} - \hat{q}_k) \\
=&\sum_{s_{-i}} \frac{\Pr(s_i = 0, s_{-i})}{\Pr(s_i = 0)} e(s_i = 0, s_{-i}) \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) (\hat{q}_{k+1} - \hat{q}_k) \\
=&\frac{1}{\Pr(s_i = 0)} \sum_{s_{-i}} h(s_i = 0, s_{-i}) \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i}) (\hat{q}_{k+1} - \hat{q}_k) \\
=&\frac{1}{\Pr(s_i = 0)} \sum_{s_{-i}} h(s_i = 0, s_{-i}) \Delta(s_{-i})
\end{aligned}
$$

where $\Delta(s_{-i}) := \sum_{k=0}^{N-1} \tilde{p}_k(s_{-i})(\hat{q}_{k+1} - \hat{q}_k)$. Since $\hat{q}_k$ is monotone, $\Delta(s_{-i})$ has the same sign for all s_{-i}. Similarly, we have

$$\Pi(1|s_i = 1) - \Pi(0|s_i = 1) = \frac{1}{\Pr(s_i = 1)} \sum_{s_{-i}} h(s_i = 1, s_{-i}) \Delta(s_{-i}).$$

Suppose $\Delta(s_{-i})$ is non-negative. We have shown earlier that $h(s_i = 1, s_{-i}) > h(s_i = 0, s_{-i})$. Thus $\sum_{s_{-i}} h(s_i = 1, s_{-i}) \Delta(s_{-i}) > \sum_{s_{-i}} h(s_i = 0, s_{-i}) \Delta(s_{-i})$ for any equilibrium that

is non-trivial. (An equilibrium is trivial if the same verdict is reached with certainty. If an equilibrium is non-trivial and the voting rule $\hat{q}_k$ is symmetric then it can be shown that for any i there is s_{-i} where $\Delta(s_{-i}) \neq 0$.) For a non-trivial equilibrium there are the following two cases:

1. If $\Pi(1|s_i = 0) - \Pi(0|s_i = 0) \geq 0$ then $\Pi(1|s_i = 1) - \Pi(0|s_i = 1) > 0$. In this cases juror i has a strict incentive to report guilty ("1") given a guilty signal ("1"). Thus $a_i = 1 \geq b_i$.

2. If $\Pi(1|s_i = 0) - \Pi(0|s_i = 0) < 0$ then the juror has strictly no incentive to report guilty given an innocent signal. Thus $a_i \geq b_i = 0$.

Observe that regardless of which case obtains, $a_i \geq b_i$. Since the argument does not depend on the identity of the juror, it follows that $a_i \geq b_i$ for all i, or in other words $a_i - b_i$ has the same sign for all i. A similar argument holds when $\Delta(s_{-i})$ is non-positive. Thus we have shown that for any non-trivial equilibrium of a voting game with a monotone rule, $a_i - b_i$ must has the same sign for all i, and hence the equilibrium payoff cannot be worse than no voting. On the other hand, a trivial equilibrium is not worse than no voting either clearly. Hence, no such voting can extract information for DM. □

III.6.14 Proof of Proposition III.8

Proof. Suppose q beats the benchmark. Claim that the following strategy profile constitutes a perfect Bayesian equilibrium in $\Gamma(q)$:

- Every expert sends the message that is the same as his signal.
- DM chooses the option that is the same as the intermediator's message.

It is clear that the strategy profile is outcome-equivalent to truthtelling under q. We now show the strategy profile is incentive compatible. For the experts, that truthtelling is incentive compatible under q immediately implies truthtelling is incentive compatible under $\Gamma(q)$ because of outcome equivalence. Consider DM. Upon receiving message "R'' the expected payoff from choosing option R is

$$\frac{1}{\sum_{s\in S} p(s)q(s)} \Big(\sum_{s\in S} p(s)d(s)q(s) \Big) > 0$$

since the term in the brackets is positive due to q beating the benchmark. Thereofre choosing R upon message "R" is best responding. If choosing S upon signal "S" is not best responding then deviating to choosing R upon signal "S" is for DM. It follows that always choosing

R regardless of the message gives DM a higher payoff than following the prescribed strategy. However, always choosing R yields a payoff no higher than the benchmark B whereas following the strategy yields a payoff higher than R given that q beats the benchmark, a contradiction. Therefore the strategy is also incentive compatible for DM. □

III.6.15 Proof of Lemma III.10

Proof. It is sufficient to show that Z represents a garbling probability function that translates (d, e, p, S) into an environment equivalent to (d', e', p', S') where $z(s'(j)|s(k)) = Z(j, k)$. That Z is non-negative and $Z^{\mathrm{T}}\mathbf{1}_{|S'|}$ imply that indeed entries in the kth column of Z constitute a probability function $z(\cdot|s(k))$. Let (d^*, e^*, p^*, S') denote the resulting environment from the information manipulation implied by Z. Thus for any $s'(j) \in S'$ we have

$$\begin{aligned} p^*(s'(j))e^*(s'(j)) &= p^*(s'(j)) \sum_{s\in S} \frac{p(s)z(s'(j)|s)}{p^*(s'(j))} e(s) \\ &= \sum_{s\in S} z(s'(j)|s)[p(s)e(s)] = Z_j\mathbf{g} \end{aligned}$$

where Z_j denotes the jth row of Z. On the other hand we have

$$Z_j\mathbf{g} = \mathbf{g}'_j = p'(s'(j))e'(s'(j))$$

by assumption. Thus $p^*(s'(j))e^*(s'(j)) = p'(s'(j))e'(s'(j))$ for every $s'(j) \in S'$. Similarly we have $p^*(s'(j))d^*(s'(j)) = p'(s'(j))d'(s'(j))$ for every $s'(j) \in S'$. Thus (d^*, e^*, p^*, S') is equivalent to (d', e', p', S'). □

III.6.16 Proof of Proposition III.9

Proof. We first show the "only if" direction. Suppose (d, e, p, S) can be translated into (d', e', p', S') by information manipulation. By Lemma III.10 there is a matrix Z such that $Z\mathbf{h} = \mathbf{h}'$ and $Z^T\mathbf{1}_{|S'|} = \mathbf{1}_{|S|}$. Thus we have

$$\sum_{s'\in S'} h'(s') = \sum_{i=1}^{|S'|}\sum_{j=1}^{|S|} Z_{ij}h(s(j)) = \sum_{j=1}^{|S|} h(s(j)) \sum_{i=1}^{|S'|} Z_{ij} = \sum_{j=1}^{|S|} h(s(j)) = \sum_{s\in S} h(s).$$

Similarly we have $\sum_{s\in S} g(s) = \sum_{s'\in S'} g'(s')$. To show part 2, denote $K := \{k = 1, ..., |S| : h(s(k)) > 0\}$ and $K' := \{k = 1, ..., |S'| : h'(s'(k)) > 0\}$. Observe that

$$\begin{aligned}\sum_{i\in K'} h'(s'(i)) &= \sum_{i\in K'}\sum_{j=1}^{|S|} Z_{ij}h(s(j)) \le \sum_{i\in K'}\sum_{j\in K} Z_{ij}h(s(j)) \\ &= \sum_{j\in K} h(s(j)) \sum_{i\in K'} Z_{ij} \le \sum_{j\in K} h(s(j)),\end{aligned}$$

which is equvalent to $\sum_{s\in S,h(s)>0} h(s) \ge \sum_{s'\in S',h'(s)>0} h'(s')$. Similarly we have $\sum_{s\in S,g(s)>0} g(s) \ge \sum_{s'\in S',g'(s')>0} g'(s')$.

Now we show the "if" direction. Suppose conditions 1 and 2 are satisfied. Construct $|S'| \times |S|$ matrix Z such that for any $i = 1, ..., |S'|$ and $j = 1, ..., |S|$:

- If $h'(s'(i)) > 0$ and $h(s(j)) > 0$:

$$Z_{ij} = \frac{h'(s'(i))}{\sum_{s\in S,h(s)>0} h(s)} + \frac{1}{|S'|}\Big(1 - \frac{\sum_{s'\in S',h'(s')>0} h'(s')}{\sum_{s\in S,h(s)>0} h(s)}\Big).$$

- If $h'(s'(i)) > 0$ and $h(s(j)) < 0$:

$$Z_{ij} = \frac{1}{|S'|}\Big(1 - \frac{\sum_{s'\in S',h'(s')<0} h'(s')}{\sum_{s\in S,h(s)<0} h(s)}\Big).$$

- If $h'(s'(i)) < 0$ and $h(s(j)) > 0$:

$$Z_{ij} = \frac{1}{|S'|}\Big(1 - \frac{\sum_{s'\in S',h'(s')>0} h'(s')}{\sum_{s\in S,h(s)>0} h(s)}\Big).$$

- If $h'(s'(i)) < 0$ and $h(s(j)) < 0$:

$$Z_{ij} = \frac{h'(s'(i))}{\sum_{s\in S,h(s)<0} h(s)} + \frac{1}{|S'|}\Big(1 - \frac{\sum_{s'\in S',h'(s')<0} h'(s')}{\sum_{s\in S,h(s)<0} h(s)}\Big).$$

Note that conditions 2 implies $\frac{\sum_{s'\in S',h'(s')>0} h'(s')}{\sum_{s\in S,h(s)>0} h(s)} \in [0, 1]$ if there exists $s \in S$ such that $h(s) > 0$. Conditions 1 and 2 imply that $\sum_{s'\in S',h'(s')<0} h'(s') > \sum_{s\in S,h(s)<0} h(s)$, which in turn implies that $\frac{\sum_{s'\in S',h'(s')<0} h'(s')}{\sum_{s\in S,h(s)<0} h(s)} \in [0, 1]$ if there exists $s \in S$ such that $h(s) < 0$. Given these observations it is straightforward to verify that Z_{ij} is non-negative for any i, j.

For any i where $h'(s'(i)) > 0$ we have

$$\begin{aligned}
\sum_{j=1}^{|S|} Z_{ij} h(s(j)) = &\sum_{j\in\{1,\ldots,|S|\}}^{h(s(j))>0} \frac{h(s(j))}{\sum_{s\in S, h(s)>0} h(s)} h'(s'(i)) \\
&+ \frac{1}{|S'|} \sum_{j\in\{1,\ldots,|S|\}}^{h(s(j))>0} \Big(h(s(j)) - \frac{\sum_{s'\in S', h'(s')>0} h'(s')}{\sum_{s\in S, h(s)>0} h(s)} h(s(j)) \Big) \\
&+ \frac{1}{|S'|} \sum_{j\in\{1,\ldots,|S|\}}^{h(s(j))<0} \Big(h(s(j)) - \frac{\sum_{s'\in S', h'(s')<0} h'(s')}{\sum_{s\in S, h(s)<0} h(s)} h(s(j)) \Big) \\
=&h'(s'(i)) + \frac{1}{|S'|} \Big(\sum_{s\in S}^{h(s)>0} h(s) - \sum_{s'\in S'}^{h'(s')>0} h'(s') \Big) \\
&+ \frac{1}{|S'|} \Big(\sum_{s\in S}^{h(s)<0} h(s) - \sum_{s'\in S'}^{h'(s')<0} h'(s') \Big) \\
=&h'(s'(i)) + \frac{1}{|S'|} \Big(\sum_{s\in S} h(s) - \sum_{s'\in S'} h'(s') \Big) \\
=&h'(s'(i)).
\end{aligned}$$

Similarly $\sum_{j=1}^{|S|} Z_{ij} h(s(j)) = h'(s'(i))$ if $h'(s'(i)) < 0$. Therefore, $Z\mathbf{h} = \mathbf{h}'$. That $Z\mathbf{g} = \mathbf{g}'$ is established analogously.

For any j where $h(s(j)) > 0$ we have

$$\sum_{i=1}^{|S'|} Z_{ij} = \sum_{i\in\{1,\ldots,|S'|\}}^{h'(s'(i))>0} \frac{h'(s'(i))}{\sum_{s\in S, h(s)>0} h(s)} + 1 - \frac{\sum_{s'\in S', h'(s')>0} h'(s')}{\sum_{s\in S, h(s)>0} h(s)} = 1.$$

Similarly $\sum_{i=1}^{|S'|} Z_{ij} = 1$ for any j where $h(s(j)) < 0$. Thus $Z^T \mathbf{1}_{|S'|} = \mathbf{1}_{|S|}$. □

III.6.17 Proof of Lemma III.7

Proof. Af first, we show that q^* gives the same expected payoff as q .

$$\begin{aligned}
&\sum_{s\in S} p(s)e(s)q^*(s)\\
=&\sum_{s\in S} p(s)e(s)\frac{1}{n!}\sum_{k=1}^{n!} q(\sigma_k(s))\\
=&\frac{1}{n!}\sum_{s\in S}\sum_{k=1}^{n!} p(\sigma_k(s))e(\sigma_k(s))q(\sigma_k(s))\\
=&\frac{1}{n!}\sum_{k=1}^{n!}\sum_{s\in S} p(\sigma_k(s))e(\sigma_k(s))q(\sigma_k(s))\\
=&\frac{1}{n!}n!\sum_{s\in S} p(s)e(s)q(s)\\
=&\sum_{s\in S} p(s)e(s)q(s)
\end{aligned}$$

Secondly, we want to show that q^* is feasible. It is straightforward to see $q^*(s) \in [0,1]$ for all s. The key step is checking ICs.

$$\begin{aligned}
&\sum_{s_{-i}} p(s_i,s_{-i})e(s_i,s_{-i})q^*(s_i,s_{-i})\\
=&\sum_{s_{-i}} p(s_i,s_{-i})e(s_i,s_{-i})\frac{1}{n!}\sum_{k=1}^{n!} q(\sigma_k(s_i,s_{-i}))\\
=&\frac{1}{n!}\sum_{s_{-i}}\sum_{k=1}^{n!} p(s_i,s_{-i})e(s_i,s_{-i})q(\sigma_k(s_i,s_{-i}))\\
=&\frac{1}{n!}\sum_{k=1}^{n!}\sum_{s_{-i}} p(\sigma_k(s_i,s_{-i}))e(\sigma_k(s_i,s_{-i}))q(\sigma_k(s_i,s_{-i}))\\
=&\frac{1}{n!}\sum_{j=1}^{n}(n-1)!\Big[\sum_{s_{-j}} p(s_j=s_i,s_{-j})e(s_j=s_i,s_{-j})q(s_j=s_i,s_{-j})\Big]\\
=&\frac{1}{n}\sum_{j=1}^{n}\sum_{s_{-j}} p(s_j=s_i,s_{-j})e(s_j=s_i,s_{-j})q(s_j=s_i,s_{-j}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\sum_{s_{-i}} p(s_i, s_{-i}) e(s_i, s_{-i}) q^*(s_i', s_{-i}) \\
=&\frac{1}{n} \sum_{j=1}^{n} \sum_{s_{-j}} p(s_j = s_i, s_{-j}) e(s_j = s_i, s_{-j}) q(s_j' = s_i', s_{-j}).
\end{aligned}$$

Thus, ICs under q^* is guaranteed by the fact q satisfies ICs. □

BIBLIOGRAPHY

Ambrus, Attila and Shih En Lu, (2014). "Almost Fully Revealing Cheap Talk with Imperfectly Informed Senders," *Games and Economic Behavior*, *88*, pp. 174 – 189, DOI: http://dx.doi.org/https://doi.org/10.1016/j.geb.2014.09.001.

Arrow, K., (1999). "Discounting, Morality, and Gaming," In P. Portney and J. Weyant (eds.) *Discounting and Intergenerational Equity*, New York, London: Resources for the Future, Chapter 2, pp. 13–22.

Asheim, G. and S. Zuber, (2014). "Escaping the Repugnant Conclusion: Rank-Discounted Utilitarianism with Variable Population," *Theoretical Economics*, *9*(3), pp. 629–650.

Azevedo, Eduardo M and Eric Budish, (2019). "Strategy-Proofness in the Large," *The Review of Economic Studies*, *86*(1), pp. 81–116.

Bain, J., (1960). "Criteria for Undertaking Water-Resource Developments," *American Economic Review*, *50*(2), pp. 310–320.

Barberà, Salvador, (2011). "Strategy-Proof Social Choice," In Kenneth J Arrow, Amartya Sen, and Kotaro Suzumura (eds.) *Handbook of Social Choice and Welfare*, *2*, Netherlands: North Holland: Elsevier, Chapter 25, pp. 731–831.

Battaglini, Marco, (2002). "Multiple Referrals and Multidimensional Cheap Talk," *Econometrica*, *70*(4), pp. 1379–1401, DOI: http://dx.doi.org/10.1111/1468-0262.00336.

Battaglini, Marco, (2004). "Policy Advice with Imperfectly Informed Agents," *4*, pp. 1100–1100.

Bergemann, Dirk and Stephen Morris, (2005). "Robust Mechanism Design," *Econometrica*, *73*(6), pp. 1771–1813.

Boadway, R., (2012). *From Optimal Tax Theory to Tax Policy: Retrospective and Prospective Views*, Cambridge, London: MIT Press.

Börgers, Tilman and Jiangtao Li, (2019). "Strategically Simple Mechanisms," *Econometrica*, *87*(6), pp. 2003–2035.

Caplin, A. and J. Leahy, (2004). "The Social Discount Rate," *Journal of Political Economy*, *112*(6), pp. 1257–1268.

Chambers, C. and T. Hayashi, (2006). "Preference Aggregation under Uncertainty: Savage vs. Pareto," *Games and Economic Behavior*, *54*(2), pp. 430–440.

Chambers, Christopher P and Federico Echenique, (2018). "On Multiple Discount Rates," *Econometrica*, *86*(4), pp. 1325–1346.

Crawford, Vincent P. and Joel Sobel, (1982). "Strategic Information Transmission," *Econometrica*, *50*(6), pp. 1431–1451.

Crémer, Jacques and Richard P McLean, (1985). "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist when Demands are Interdependent," *Econometrica*, *53*, pp. 345–361.

Crémer, Jacques and Richard P. McLean, (1988). "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, *56*(6), pp. 1247–1257.

De Meyer, B. and P. Mongin, (1995). "A Note on Affine Aggregation," *Economics Letters*, *47*(2), pp. 177–183.

Dhillon, A. and J.-F. Mertens, (1999). "Relative Utilitarianism," *Econometrica*, *67*(3), pp. 471–498.

Drugeon, J.-P. and B. Wigniolle, (2016). "On Time-Consistent Policy Rules for Heterogeneous Discounting Programs," *Journal of Mathematical Economics*, *63*, pp. 174–187.

Drugeon, J.-P. and B. Wigniolle, (2017). "On Time-Consistent Collective Choice with Heterogeneous Quasi-Hyperbolic Discounting," Working Paper, Paris School of Economics.

Eckstein, O., (1957). "Investment Criteria for Economic Development and the Theory of Intertemporal Welfare Economics," *Quarterly Journal of Economics*, *71*(1), pp. 56–85.

Farhi, E. and I. Werning, (2007). "Inequality and Social Discounting," *Journal of Political Economy*, *115*(3), pp. 365–402.

Feldstein, M., (1964). "The Social Time Preference Discount Rate in Cost Benefit Analysis," *Economic Journal*, *74*(294), pp. 360–379.

Feng, Tangren and Shaowei Ke, (2018). "Social Discounting and Intergenerational Pareto," *Econometrica*, *86*(5), pp. 1537–1567.

Feng, Tangren and Qinggong Wu, (2019). "Getting Information from Your Enemies," Working Paper, University of Michigan.

Feng, Tangren and Qinggong Wu, (2020). "Robust Binary Voting," Working Paper, University of Michigan.

Fishburn, P., (1984). "On Harsanyi's Utilitarian Cardinal Welfare Theorem," *Theory and Decision*, *17*(1), pp. 21–28.

Fleurbaey, M. and S. Zuber, (2015). "Discounting, Risk and Inequality: A General Approach," *Journal of Public Economics*, *128*, pp. 34–49.

Frederick, S., G. Loewenstein, and T. O'Donoghue, (2002). "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, *40*(2), pp. 351–401.

Galperti, S. and B. Strulovici, (2017). "A Theory of Intergenerational Altruism," *Econometrica*, *85*(4), pp. 1175–1218.

Gerardi, Dino, Richard McLean, and Andrew Postlewaite, (2009). "Aggregation of Expert Opinions," *Games and Economic Behavior*, *65*(2), pp. 339 – 371, DOI: http://dx.doi.org/https://doi.org/10.1016/j.geb.2008.02.010.

Gershkov, Alex, Benny Moldovanu, and Xianwen Shi, (2016). "Optimal Voting Rules," *The Review of Economic Studies*, *84*(2), pp. 688–717.

Gibbard, Allan, (1973). "Manipulation of Voting Schemes: a General Result," *Econometrica*, *41*(4), pp. 587–601.

Gilligan, Thomas W. and Keith Krehbiel, (1989). "Asymmetric Information and Legislative Rules with a Heterogeneous Committee," *American Journal of Political Science*, *33*(2), pp. 459–490.

Gollier, C. and R. Zeckhauser, (2005). "Aggregation of Heterogeneous Time Preferences," *Journal of Political Economy*, *113*(4), pp. 878–896.

Halevy, Y., (2015). "Time Consistency: Stationarity and Time Invariance," *Econometrica*, *83*(1), pp. 335–352.

Hammond, P., (1987). "Altruism," In J. Eatwell, M. Milgate, and P. Newman (eds.) *The New Palgrave: A Dictionary of Economics*, Basingstoke: Palgrave Macmillan, 1st edition, pp. 85–87.

Harsanyi, J., (1955). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, *63*(4), pp. 309–321.

Harsanyi, J, (1967/1968). "Games with Incomplete Information Played by Bayesian Agents i-iii," *Management Science*, *14*(159–182, 320–334, 486–502).

Hylland, Aanund, (1980). "Strategy Proofness of Voting Procedures with Lotteries as Outcomes and Infinite Sets of Strategies," *Unpublished paper, University of Oslo*.

Jackson, M. and L. Yariv, (2014). "Present Bias and Collective Dynamic Choice in the Lab," *American Economic Review*, *104*(12), pp. 4184–4204.

Jackson, M. and L. Yariv, (2015). "Collective Dynamic Choice: The Necessity of Time Inconsistency," *American Economic Journal: Microeconomics*, *7*(4), pp. 150–178.

Jehiel, Philippe, Moritz Meyer-ter Vehn, Benny Moldovanu, and William R Zame, (2006). "The Limits of Ex Post Implementation," *Econometrica*, *74*(3), pp. 585–610.

Jonsson, A. and M. Voorneveld, (2018). "The Limit of Discounted Utilitarianism," *Theoretical Economics*, *13*(1), pp. 19–37.

Karni, E., (1998). "Impartiality: Definition and Representation," *Econometrica*, *66*(6), pp. 1405–1415.

Krishna, Vijay and John Morgan, (2001). "Asymmetric Information and Legislative Rules: Some Amendments," *The American Political Science Review*, *95*(2), pp. 435–452.

Laibson, D., (1997). "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, *112*(2), pp. 443–478.

Marglin, S., (1963). "The Social Rate of Discount and the Optimal Rate of Investment," *Quarterly Journal of Economics*, *77*(1), pp. 95–111.

Mertens, Jean-François and Shmuel Zamir, (1985). "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, *14*(1), pp. 1–29.

Millner, A. and G. Heal, (2018). "Time Consistency and Time Invariance in Collective Intertemporal Choice," *Journal of Economic Theory*, *176*, pp. 158–169.

Millner, Antony, (2020). "Nondogmatic Social Discounting," *American Economic Review*, *110*(3), pp. 760–75.

Mongin, P., (1995). "Consistent Bayesian Aggregation," *Journal of Economic Theory*, *66*(2), pp. 313–351.

Mongin, P., (1998). "The Paradox of the Bayesian Experts and State-Dependent Utility Theory," *Journal of Mathematical Economics*, *29*(3), pp. 331–361.

Moulin, Hervé, (1980). "On Strategy-Proofness and Single Peakedness," *Public Choice*, *35*(4), pp. 437–455.

Nordhaus, W., (2007). "A Review of the Stern Review on the Economics of Climate Change," *Journal of Economic Literature*, *45*(3), pp. 686–702.

Piacquadio, P., (2017). "The Ethics of Intergenerational Risk," Working Paper, University of Oslo.

Pigou, A., (1920). *The Economics of Welfare*, London: Macmillan & Co., Limited.

Ramsey, F., (1928). "A Mathematical Theory of Saving," *Economic Journal*, *38*(152), pp. 543–559.

Ray, D., N. Vellodi, and R. Wang, (2017). "Backward discounting," Working Paper, New York University.

Satterthwaite, Mark Allen, (1975). "Strategy-Proofness and Arrow's conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions," *Journal of Economic Theory*, *10*(2), pp. 187–217.

Segal, U., (2000). "Let's Agree That All Dictatorships Are Equally Bad," *Journal of Political Economy*, *108*(3), pp. 569–589.

Sen, A., (1961). "On Optimizing the Rate of Saving," *Economic Journal*, *71*(283), pp. 479–496.

Solow, R., (1974). "The Economics of Resources or the Resources of Economics," *American Economic Review*, *64*(2), pp. 1–14.

Stern, N., (2007). *The Economics of Climate Change: The Stern Review*, Cambridge, UK: Cambridge University Press.

Strotz, R., (1955). "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, *23*(3), pp. 165–180.

Weitzman, M., (2001). "Gamma Discounting," *American Economic Review*, *91*(1), pp. 260–271.

Wolinsky, Asher, (2002). "Eliciting Information from Multiple Experts," *Games and Economic Behavior*, *41*(1), pp. 141 – 160, DOI: `http://dx.doi.org/https://doi.org/10.1016/S0899-8256(02)00003-9`.

Zuber, S., (2011). "Can Social Preferences Be Both Stationary and Paretian?" *Annals of Economics and Statistics*(101/102), pp. 347–360.

Zuber, S. and G. Asheim, (2012). "Justifying Social Discounting: The Rank-Discounted Utilitarian Approach," *Journal of Economic Theory*, *147*(4), pp. 1572–1601.